THESIS FOR THE DEGREE OF DOCTOR OF ENGINEERING

Statistical assessment of genomic variability in tumours and bacterial communities

Anna Rehammar

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2019

Statistical assessment of genomic variability in tumours and bacterial communities
Anna Rehammar
Göteborg, 2019
ISBN 978-91-7905-135-8

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4602
ISSN 0346-718X

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX.
Printed in Chalmers Reproservice Göteborg, 2019

Till Britt-Marie
Jag saknar dig

Statistical assessment of genomic variability in tumours and bacterial communities

Anna Rehammar

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Current high-throughput DNA sequencing technologies have the ability to generate large amounts of high-resolution genomic data. The high dimensionality in combination with the substantial levels of technical errors and biological variability typically present in the data make, however, the analysis challenging. Tailored statistical methods are therefore crucial for reaching valid biological conclusions. In this thesis, such methods were developed and applied to address research questions in biology and medicine.

First, a method for identification of tumour-specific (somatic) mutations was developed, which included steps for noise-reduction, sensitive detection of DNA alterations and removal of systematic errors. In Paper I, the method was applied to exome-sequenced paired normal–tumour samples from pheochromocytoma patients. A significantly higher mutation rate was found in malignant compared to benign tumours and three genes with recurrent somatic mutations, exclusively located in malignant tumours, were identified. In paper II and III, somatic mutations were identified in patients with acute myeloid leukemia and evaluated as biomarkers in personalised deep sequencing analysis of remaining cancer cells after treatment. In paper III, a statistical model correcting for position-specific errors in the data was developed and shown to provide superior sensitivity compared to standard techniques. In paper IV, clinically relevant molecular subgroups of metastatic small intestinal neuroendocrine tumours were identified based on miRNA gene expression profiles. Survival analysis and subsequent validation suggested miR-375 as a prognostic biomarker. In paper V, a hierarchical Bayesian model for detecting differences on nucleotide level between microbial communities is proposed. By including between-sample variability and utilizing a shrinkage approach, the model was able to perform well both in cases of few samples and high biological variability. Finally, the model was used to detect antibiotic resistance mutations in bacteria.

This thesis demonstrates that dedicated statistical analysis and knowledge of the underlying error structure present in high-dimensional biological data is of importance for enabling accurate interpretation and sound conclusions.

Keywords: high-throughput sequencing, somatic mutations, cancer genetics, personalised diagnostics, metagenomics, hierarchical Bayesian modelling

List of publications

The thesis includes the following papers:

I. Wilzén, A*, **Rehammar, A***, Muth, A, Nilsson, O, Tesan Tomic, T, Wängberg, B, Kristiansson, E, Abel, F (2016). Malignant pheochromocytomas/paragangliomas harbor mutations in transport and cell adhesion genes. *International Journal of Cancer.* 138(9).

II. Malmberg, E*, Ståhlman, S*, **Rehammar, A**, Samuelsson, T, Alm, SJ, Kristiansson, E, Abrahamsson, J, Garelius, H, Palmqvist, L, Fogelstrand, L (2017). Patient-tailored analysis of minimal residual disease in acute myeloid leukemia using next generation sequencing. *European Journal of Haematology.* 98(1).

III. Malmberg, ED, **Rehammar, A***, Pereira, MB*, Abrahamsson, J, Samuelsson, T, Ståhlman, S, Asp, J, Tierens, A, Palmqvist, L, Kristiansson, E, Fogelstrand, L (2018). Accurate and Sensitive Analysis of Minimal Residual Disease in Acute Myeloid Leukemia Using Deep Sequencing of Single Nucleotide Variations. *Journal of Molecular Diagnostics.* 21(1).

IV. Arvidsson, Y, **Rehammar, A**, Bergström, A, Andersson, E, Altiparmak, G, Swärd, C, Wängberg, B, Kristiansson, E, Nilsson, O (2018). miRNA profiling of small intestinal neuroendocrine tumors defines novel molecular subtypes and identifies miR-375 as a biomarker of patient survival. *Modern Pathology.* 31(8).

V. **Rehammar, A.** and Kristiansson, E. (2019). A hierarchical Bayesian model for assessing differential nucleotide composition between metagenomes. *Manuscript.*

Additional papers not included in this thesis:

VI. Tesan Tomic, T*, Olausson, J*, **Rehammar, A**, Deland, L, Ejeskär, K, Nilsson, S, Kristiansson, E, Muth, A, Wängberg, B, Nilsson, O, Abel, F (2019). MYO5B mutations in pheochromocytoma/paraganglioma tumors promote cancer progression. *Submitted.*

VII. Hofving, T, Elias, E, Inge, L, Altiparmak, G, **Rehammar, A.**, Kristiansson, E, Nilsson, O*, Arvidsson, Y* (2019). SMAD4 haploinsufficiency in small intestinal neuroendocrine tumours. *Manuscript.*

VIII. Bengtsson-Palme, J, Boulund, F, Edström, R, Feizi, A, Johnning, A, Jonsson, VA, Karlsson, FH, Pal, C, Pereira, MB, **Rehammar, A**, Sanchez, J, Sanli, K, Thorell, K (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *PROTEOMICS* 16(18).

* Authors contributed equally

Author contributions

I. Performed the bioinformatical work and statistical analyses, including preprocessing of the data, calling of germline and somatic mutations, annotation, filtering, removal of likely artifacts and gene enrichment analysis. Participated in study design, interpretation of the results and in drafting and editing the manuscript.

II. Performed the bioinformatical work and the statistical analyses for identification of somatic mutations. Developed and performed the statistical analysis for choosing MRD candidates. Participated in drafting and editing the manuscript.

III. Performed the bioinformatical work and the statistical analyses for identification of somatic mutations. Developed the statistical model for estimation of variant allele frequency in deep sequencing data. Performed the statistical analyses and participated in study design and interpretation of the results. Participated in drafting and editing the manuscript.

IV. Performed the analysis of gene expression measured by miRNA microarrays and tissue microarrays, including quality assessment, background correction and normalization of the miRNA microarrays, identification of differentially expressed genes and survival analysis. Performed the clustering and statistical tests of association. Participated in drafting and editing the manuscript.

V. Designed the study, performed the bioinformatical work with the data used for resampling, developed the statistical model, and implemented the model and the simulations. Performed the analyses and interpreted the results. Drafted and edited the manuscript.

Acknowledgements

I am happy to get to thank all the nice and competent people I have been working with during the past years. First and foremost, I want to express my deepest thanks to my supervisor Erik Kristiansson. You have an outstanding ability to share your enthusiasm for research that have had a large impact on both me and the atmosphere in our group. I have really appreciated your way of supervising me, with trust to work independently but with full support whenever I asked for it. You are one of the most non-hierarchical and prestigeless persons I have met, without taking one step down from the responsibilities as a leader and expert. Your constructive view and always so positive attitude have been indispensable both in my research and when life takes unexpected turns. I am impressed by your way of sticking to your core values and express them with clarity.

I would also like to thank my co-supervisor at Sahlgrenska Academy, Frida Abel, for sharing your curiosity about elucidating the cancer pathways and for supporting me. To my examiners and (co-)supervisors through the years of master and doctoral studies at Mathematical Sciences – Olle Nerman, Staffan Nilsson, Aila Särkkä, Marita Olsson and Petter Mostad – it has been a privilege to have you around, both for your great knowledge and your kind hearts. You all make academia a nicer and more humane place. Thanks also to all the friendly and dedicated seniors at the department that have taken on the task to teach me statistics in courses and discussions – Kerstin, Jacques, Serik, Torgny, Olle H, Rebecka, Ziad and Dragi. My appreciation also goes to the always so helping colleagues Lotta, Johan, Marianne, Marie and many more.

I am thankful to all my co-authors within biology and medicine for the opportunity to work together on important questions. To Erik Delsing Malmberg, Linda Fogelstrand, Yvonne Arvidsson and Ola Nilsson Wassén, I have really appreciated our close collaborations with in-depth discussions, knowledge-sharing, positive atmosphere and your trust in me to analyse the data. As always, special thanks to Annica Wilzén, I truly miss you in my daily research life but I am grateful for every minute we still get to talk and laugh together. I am proud of the work we did together! I also want to thank Marcela Dávila López for your helping attitude and keeping the Gothenburg bioinformatics network alive.

Thank you Viktor Jonsson for being both the best room mate and a friend. I am impressed by your ability to reflect and grateful that you want to share your thoughts with me. Anders Sjögren, thanks for your help with statistical questions and for being someone wise and warm to share life moments with. And of course, a big thanks to all the present and former members of Erik Kristiansson's research group. Fanny Berglund, Anna Johnning, Mikael Gustavsson, Tobias Österlund, Fredrik Boulund, Mariana Pereira, Johannes Dröge

and Johan Bengtsson-Palme – you are all awesome. Life at the department would not have been the same without you. Special thanks to Mariana for nice partnership in the leukaemia project.

An important part of work is morning chats and nice company at lunches and fika. Thanks to all my colleagues at the department, and especially Malin, Claes, Marina, Johan S, Henrik, Oskar A, Jan, Olle E, Jonathan K, Jonathan N, Ivar, Maria, Patrik, Anton, Olof, Juan, Helga, Sandra and Tobias A. This was equally important back in 2006/2007 when I got to know Alexandra, Sofia, Janeli and Oscar H – thank you for introducing me to life in academia.

Thanks to all my relatives and friends for supporting me in so many different ways. I have really needed all those long walks! To my parents; you are the kindest persons I know and I am grateful that you care deeply not only for me, but also for Robert and our kids without ever hesitating.

Tack Robert, min älskade livskamrat, för ditt stöd under alla våra år tillsammans och för att du tycker min forskning är så cool. Till våra fantastiska barn Alvin, Vilhelm och Sofia, som signifikant förlängt min doktorandtid ($p \ll 0.05$), ni visar mig dagligen hur härligt livet kan vara. Tack för er glädje, nyfikenhet, omtanke, spring i benen och klokhet.

Anna Rehammar

Contents

Chapter 1

Introduction

In all cells of all organisms, the genetic material contains information regarding how the cells should develop and function. Parts of the genetic material are partitioned into genes, i.e. units that hold information about how other molecules, such as proteins, should be build. All the genetic material in a cell is collectively called the genome. Differences in the genome determine, together with the encountered environment, our different traits, development and responses. The evolution of new functions and organisms is possible due to changes in the genome. The genetic material consists of DNA molecules, each constructed of a long chain of four different building blocks. These are collectively called nucleotides and are denoted A, C, G and T. To permit analysis of how the information encoded in the DNA molecules govern biological processes, the information must be read. That is, the exact order of the nucleotides along the DNA molecule needs to be determined. The term "sequencing" refers to this process.

Until recently, sequencing was a time-consuming and costly task. For example, when the first human genome was sequenced, it was a large collaborative project that required more than 10 years to complete (Lander et al., 2001). An early strategy for investigating the association of a property with variations in the human genome was, therefore, to only read a very limited set of positions instead of the whole sequence. However, new innovative techniques for DNA sequencing, commonly referred to as the next-generation sequencing (NGS), have dramatically lowered the cost and efforts, and revolutionised the ability to characterise genomes (Mardis, 2011). It is now possible to compare information for the whole human genome, or substantial parts of it, from many samples. The relation between genetic alterations and different phenotypic properties, such as, for example, diseases, can thereby be investigated at an unprecedented resolution.

An impressive example of what is now possible is the whole-genome se-

quencing of in total 15,220 Icelanders, where the data for the first 2,636 individuals were reported in 2015 and then updated to the current set in 2017 (Gudbjartsson et al., 2015; Jónsson et al., 2017). The data is paired with other unique resources for the Icelandic population, such as a genealogy for the nation documented several hundred years prior, access to nationwide healthcare information and additional DNA sequence data for more than 150,000 Icelanders previously analysed at a lower resolution. In Gudbjartsson et al. (2015) the landscape of genetic variants in the human genome in relation to, for example, functional annotation and gene position is described. Furthermore, three examples of connections between genetic variants and diseases found using the data are presented, and additional such findings have been reported in subsequent articles, see for example Oddsson et al. (2015) and Haraldsdottir et al. (2017). Even larger projects are ongoing, reflecting that availability of genome sequencing data will continue to increase (Turnbull et al., 2018).

The alterations in the nucleotide sequences that give rise to the genetic variability are called mutations. The different genetic variants that mutations create are called alleles. Mutations can occur in several different ways. As a first example, there can be an exchange (also called substitution) of one nucleotide for another, which will herein be denoted a single nucleotide variant (SNV). Furthermore, one or a few nucleotides can be inserted or deleted from the DNA chain, and such mutations are called insertions and deletions, respectively, or simply indels. These small-scale mutations are a focus in the work described in this thesis. A well-known example of such mutations are SNVs and indels in the *BRCA1* gene, changing the properties of the encoded protein and leading to an increased risk for breast cancer (King et al., 2003). However, there are also mutations on a larger scale, with the amplification or loss of larger regions up to whole chromosomes (a whole DNA molecule) or structural rearrangements within or between chromosomes. An example is the gain of an extra copy of chromosome number 21 or parts of it, resulting in Downs syndrome in the carrier. Although mutations can have damaging effects, it is important to remember that they are a prerequisite for evolution and the gain of new beneficial properties. One example is a mutation in the *FUT2* gene that gives rise to resistance against winter vomiting disease (Thorven et al., 2005).

1.1 Cancer genetics

In cancer, a number of mutations alter the normal functions of a cell and turn it into a cancer cell with enhanced ability to, for example, grow, divide, invade other tissue and resist cell death (Hanahan and Weinberg, 2011). The different cancer types constitute a heterogeneous group of diseases that can have large differences in their genetic causes. Even within a specific type of cancer, such as breast or lung cancer, there are many different combinations of mutations that

can give rise to a tumour (Vogelstein et al., 2013). The mutations harboured by a specific tumour influence, for example, the aggressiveness of the disease and the ability of the tumour to metastasise, i.e. to spread to other parts of the body (Armaghany et al., 2012; Brodeur et al., 1984). Furthermore, the response to treatment can be dependent on which mutations that are present in the tumour, and subsequent mutations can give rise to drug resistance during treatment (Garnett et al., 2012; Zahreddine and Borden, 2013; Nilsson et al., 2009). It is therefore important to characterise which mutations that cause different types of cancer and how they influence the progression and properties of the disease. This is required both to gain a more thorough understanding of tumour biology and to be able to develop better diagnostics and treatment.

However, the analysis regarding functional impact of mutations are impeded by the fact that tumour cells have a high mutation rate. Many of the mutations acquired during tumour growth are so called passenger mutations and do not influence the progression of the disease (Martincorena and Campbell, 2015). Additionally, a mix of inherited mutations, that exist in all cells of an individual, and acquired mutations often together cause the tumour development (Knudson, 1971).

To be able to utilise the knowledge about cancer mutations, and employ it in a personalised cancer therapy, it is important to develop the use of mutations as biomarkers and improve the techniques for identifying mutations in clinical care. Here, it is often necessary to have a capability of detecting very low amounts of cancer cells, and specific mutations in low proportions, to choose an appropriate treatment as early as possible (Shin et al., 2017; Fiala and Diamandis, 2018) .

In a cancer cell, the alterations in the genome lead, together with stimuli from the surrounding, to a number of molecular changes, e.g in the level of RNA molecules and in proteins. To study these changes, that mediates the altered information in the genome, is important to fully understand the processes in cancer cells. The information in the DNA molecules is transcribed gene-wise to RNA molecules of which some are translated to proteins. Proteins are the molecules that promote and provide control of the chemical reactions in the cell. The number of messenger RNA (mRNA) molecules transcribed from a specific gene relates to the amount of the corresponding protein. To examine changes in the levels of different mRNA molecules is a way of study the effects of mutations and the state of the cell, and the levels can also be used as biomarkers. For example, an mRNA profile that strongly predicts a short time interval until developing metastases have been identified in tumours of breast cancer patients (van't Veer et al., 2002).

Large scale quantification of mRNA levels can be accomplished by microarray analysis. It is a technique that has been extensively used for more than 20 years, where the levels of mRNA for, in principle, all genes can be determined

simultaneous by hybridising mRNA molecules to immobilised probes (Katagiri and Glazebrook, 2009). With the advent of NGS, techniques for sequencing RNA molecules and quantifying their levels have also been developed. RNA sequencing allows for a wider dynamic range and detection of novel mRNA molecules, such as fusion genes (Kukurba and Montgomery, 2015). The latter refers to novel combinations of two genes, which can arise due to structural rearrangements in the genome, and such mutations can be strong drivers of cancer (Gao et al., 2018).

1.2 Genetic variability in metagenomes

Genetic variability is also studied in microbiology. Microorganisms are vital parts of all ecosystems and organised in communities. These have historically been difficult to study due to their complexity. In particular, the methods have been dependent on the ability to culture the studied organisms in a laboratory. A microbial community can consist of thousands of species, and only a limited number of those are straight-forward to cultivate using standard protocols. However, with the advent of NGS techniques, the field of metagenomics has gained popularity. In metagenomics, all the genetic material from a sample taken directly from the environment is sequenced, without any prior cultivation. Thereby the genetic variability, and hence the compositions of species and biological functions, and its connection to different conditions and properties can be investigated.

A better understanding of the processes in microbial communities is of great importance in many different fields, such as agriculture, waste water treatment and medicine. For example, bacteria exist practically everywhere, both in the environment and within humans. They often contribute to important functions, such as the digestion process in the gut. However, bacteria can also cause infections, and we are dependent on having antibiotics to treat those infections. An emerging problem is bacteria that have become resistant to one or several types of antibiotics. This phenomenon cause infections that are today easily treatable to become life-threatening ones and can, in the long run, hamper many modern medical procedures. For example, effective antibiotics are important when performing surgery to prevent infections in wounds. Resistance towards antibiotics typically depends on changes in the genetic material of the bacteria. Mutations in protein coding genes is one way of acquiring resistance. For example, only three SNVs in the genome of the bacterium *Escherichia coli* is enough to make it highly resistant to certain types of antibiotics (Bagel et al., 1999). To be able to advance our understanding of the mechanisms underlying antibiotic resistance, one important component is thus to examine which mutations that exist in bacteria and that are promoted under selection pressure from antibiotics.

1.3 Sequencing data and the statistical challenges

As described in previous sections, the new sequencing techniques have made it possible to generate massive amounts of data and opened up a wealth of new opportunities for analysis of genomes. There are, however, also a number of new challenges related to data analysis and handling. The data is high-dimensional, both in the sense that it contains information at nucleotide resolution for the genetic material, and as a result of the new way the data is generated. In addition, the error profile of the data is fundamentally different from earlier techniques and requires careful bioinformatical handling and statistical evaluation for adequate interpretation. (Mardis, 2011).

In general, NGS data can be generated in large amounts, but consists typically of small pieces with overlapping and noisy information that needs to be concatenated and evaluated to reach consensus. After extraction of the genetic material from the studied sample, the DNA molecules are heavily fragmented and many such fragments are then sequenced rapidly and in parallel (Metzker, 2010). The term "massively parallel sequencing" is hence often used to describe these techniques. By sequencing of millions, or even billions, of DNA fragments from multiple cells each region of interest in the genome can be covered several times. The reads (i.e., sequenced fragments) are then mapped to a reference genome, generating piles of reads (Figure 1.1).

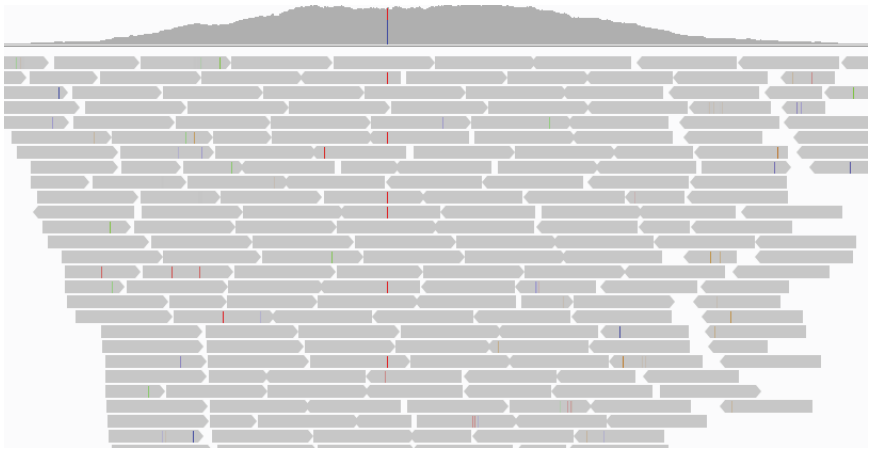

Figure 1.1: Sequenced DNA fragments are mapped to the human reference genome and viewed by the visualisation tool Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013). The coloured vertical lines represent positions where there are discrepancies (variant alleles) compared with the reference. On top, a histogram shows the number of times each position is read.

When identifying mutations using NGS data, the task is to decide for which

positions there actually are mutations and for which positions discrepancies from the reference genome only represent errors in the data. The methods need to be sensitive to detect mutations also in regions with few sequenced fragments and mutations that potentially only exist in parts of the analysed cells. At the same time, they need to have a low false positive rate, due to the high dimensionality of the data with many positions to consider. Sequencing the set of protein-coding genes in humans provides approximately 50 million positions, and the whole human genome has more than 3 billion positions.

To achieve both high sensitivity and specificity is a non-trivial task, since the data contains considerable levels of noise. These are due to errors introduced during sample preparations and sequencing of the DNA, and to limitations of the bioinformatical data processing (Olson et al., 2015). As an example, the DNA fragments are amplified, i.e. copied, in the sample preparation, which can lead to the insertion of incorrect nucleotides and bias in what parts of the genome that are represented in the sample (Aird et al., 2011). Errors introduced during sequencing are both random and systematic. For example, errors occur more often at the ends of reads and in specific patterns in the nucleotide sequence (Minoche et al., 2011). If not accounted for, these errors can lead to biased results. Furthermore, it can be difficult to determine where a specific read should be placed along the reference genome, due to that the reads are typically short (100 nucleotides is common) and contain multiple sequencing errors. The problem is more severe for reads originating from regions with repetitive patterns and from genes that are evolutionarily closely related and hence may have similar nucleotide sequences (Treangen and Salzberg, 2011). An accumulation of incorrectly placed reads can lead to discrepancies from the reference genome that are artificial but look like true mutations.

Thus, to be able to employ NGS data and transform it into accurate information that can be used for new biological insights, knowledge about the data structure and performing a sound analysis utilising bioinformatical and statistical methods that properly handle the variability in the data, is crucial. New and tailored computational and statistical methods are therefore vital to take full advantage of the information present in the data and to reach correct biological and medical interpretations.

Chapter 2

Aims

In the papers included in this thesis, biological data produced by high-throughput techniques is utilised to assess genomic changes in cancer and bacterial communities. In all cases, the data contains a high level of variability, stemming from both biological and technical factors. The overall aim of this thesis is to apply and develop bioinformatical and statistical methods that take the high dimensionality and the variability of the data into account. In particular, the objectives are to reduce the noise, model the variability and remove systematic biases. This is vital to perform sensitive analyses and keep the false positive rate low, and thereby be able to reach valid biological conclusions. The specific aims are as follows:

1. Develop a method for identification of somatic mutations in high-throughput DNA sequence data. The method should be sensitive, have a low false positive rate and include all the steps in the analysis starting from the raw sequencing data and arriving at a set of annotated and carefully assessed somatic mutations (Paper I-II).
2. Apply and adapt the developed method (Aim 1) to data from the cancers pheochromocytoma/paraganglioma and acute myeloid leukaemia, in order to find somatic mutations that are potentially relevant for tumour development or as biomarkers (Paper I-III).
3. Develop a statistical model for quantifying low frequency mutations in targeted deep sequencing data, in order to facilitate personalised diagnostics of leukaemia (Paper III).
4. Characterise the miRNA profile of small intestinal neuroendocrine tumours to search for molecular subgroups of clinical relevance and biomarkers for tumour development and outcome (Paper IV).
5. Develop a statistical model for the detection of differences at the nucleotide level between groups of metagenomes, sampled from bacterial communities encountering different experimental conditions (Paper V).

Chapter 3

Finding somatic mutations in exome sequencing data

In Paper I and Paper II, the focus is on finding tumour-specific mutations in protein-coding regions (the exome) for the cancer types pheochromocytoma/paraganglioma and acute myeloid leukemia, respectively. To complement the relatively brief methods sections in these papers, the bioinformatical and statistical approaches used to pre-process the raw sequencing data, identify candidate somatic mutations and filter for technical valid and biological important somatic mutations are described in the sections below.

In a tumour cell, there is a mix of inherited (germline) mutations and mutations that are specific to the tumour cells, denoted somatic mutations. The search for somatic mutations is, in important aspects, different from identifying germline mutations. In particular, the ratio of sequenced fragments where the mutation is expected to exist can differ substantially. In human cells, all but the sex chromosomes are inherited pairwise, resulting in two copies of each gene. In general, all the cells in the body contain the same genetic material. On average, a germline mutation is therefore expected to show up in 50% (heterozygous) or 100% (homozygous) of the sequenced fragments. That is, the variant allele frequency (VAF) is expected to be 50% or 100%. However, in samples from tumour cells this may no longer be the case. During development and growth of a tumour, new somatic mutations are typically acquired. When a new somatic mutation appears, the mutation event occurs in one of the tumour cells. How successful that cell is in surviving and dividing into additional cells with the same mutational pattern is dependent on its selective advantage. tumours are therefore often heterogeneous and consist of multiple subclones, meaning that the genetic material differs among the cells. While some somatic mutations can be common to all cells, due to an early mutation event or a large

selective advantage, others exist only in subclones. Both types are important to find to fully understand the genetic cause of how a tumour originates, develops and responds to treatment. Furthermore, samples from tumours often contain normal cells to a certain extent. Tumour cells can also mutate to have more or less than two copies of each gene. Altogether, this means that in a tumour sample, the frequency of the variant allele for a somatic mutation can take on values from, in principle, just above 0% to 100%. Thus, the assumption concerning which VAF:s to expect in the data from a tumour sample must be relaxed compared with those employed when searching for germline mutations in a normal sample.

Another aspect is directly connected to the definition of a somatic mutation: it should not be present in normal cells. A paired experimental design, including samples from both normal and tumour cells from each patient, is therefore needed. For each position, the tumour and normal sequence data is compared and if variant alleles are present in the tumour sample but not in the normal sample, a candidate position for a somatic mutation is, in principle, found (Figure 3.1).

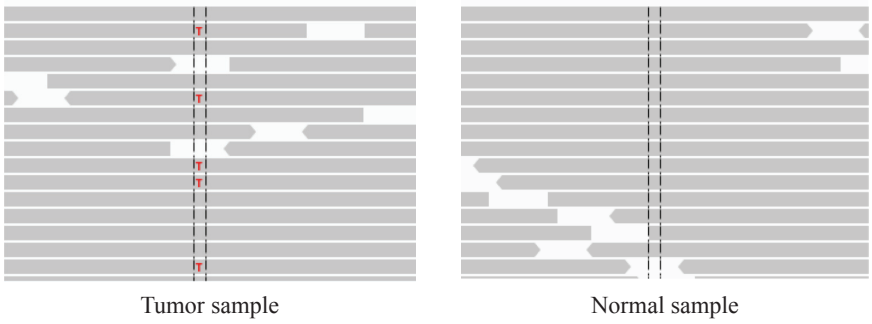

Figure 3.1: Sequenced DNA fragments aligned to the human reference genome and viewed using the Integrative Genomics Viewer, where variant alleles are shown by coloured letters. Variant alleles are detected in the tumour sample but not in the normal sample, i.e., a candidate somatic mutation.

3.1 Pre-processing of the data

The purpose of pre-processing the data, before the actual identification of the mutations, is to correct or at least compensate for errors introduced during sample preparation, sequencing and mapping to the reference genome.

We start from the point where we have access to the reads, i.e., sequenced DNA fragments, and quality scores for each sequenced nucleotide (denoted Q). The quality scores are related to the probability of a sequencing error, P,

according to

$$Q = -10 \log_{10} P,$$

where the probability values are estimated during sequencing based on a range of features for the detected signals. The first step in the preprocessing is to filter the data based on the quality scores to ensure that reads with overall low quality are discarded. Additionally, during Illumina sequencing, the quality often drops towards the end of the reads, and such stretches can be trimmed off during the filtering step (Minoche et al., 2011). In Paper I-III, we used the tool PRINSEQ to perform quality filtering and trimming (Schmieder and Edwards, 2011).

Then, the reads are mapped (aligned) to the reference genome. That is, the original position in the genome for each sequenced DNA fragment is searched. A number of different algorithms for aligning reads have been developed (Li and Homer, 2010). In paper I-III, we used the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010) in paired-end mode, as recommended in the Best Practices developed at the Broad Institute (Van der Auwera et al., 2013). Paired-end refers to the type of sequencing performed, where the DNA fragment size is aimed at being at least twice the length of a read and then the fragment is sequenced from both edges. In this way, the mapping accuracy increases, since information from both reads in a pair can be utilised. A mapping quality score is assigned to each read, indicating how well the read matched the reference sequence. Reads matching several intervals in the reference equally well are flagged by giving them a mapping quality score of zero.

The DNA amplification used in the sample preparation can lead to the same original fragment being sequenced twice or more, especially when a small amount of DNA is used as the input material. To avoid accounting for the same information several times, such duplicated reads must be removed. We used a tool called Picard (http://broadinstitute.github.io/picard) to compare read-pairs and mark those with the same genomic starting positions for both reads as duplicates. In the duplicate marking, all reads from the same sample preparation must considered simultaneously. We noted that when using formalin-fixed paraffin-embedded (FFPE) tumour material in Paper I, the levels of duplicates were generally much higher than for fresh-frozen (SF) material. FFPE samples typically have more fragmented DNA than SF samples as well as artificially introduced nucleotide changes (Do and Dobrovic, 2015). The higher duplicate levels were likely due to the additional rounds of amplification that were needed in the sample preparation of the FFPE material. Additionally, the Picard algorithm to a larger extent left duplicated reads unmarked in the data from FFPE samples, due to inconsistent mapping of one of the reads in the pair. This produced a significant amount of false positive somatic mutations in the FFPE material. We removed these by adding a down-stream filter requiring each mutation to be found in several different positions in its supporting

reads. This problem was also noticed and solved similarly in another study utilising FFPE material (Yost et al., 2012).

In regions with insertion or deletions in the sequenced sample, the mapping algorithms often encounter difficulty determining whether to include indels or mismatched nucleotides in the alignment, especially at the ends of reads. Each read-pair is mapped independently of the others, which can produce inconsistent decisions for different reads at the same position. The process of correcting for such inconsistencies is called indel realignment. Intervals that need to be corrected are searched for, and all reads in such an interval are realigned together (DePristo et al., 2011). For a paired design with samples from normal and tumour cells, it is important to perform indel realignment with all reads from one patient included at the same time. Otherwise, different consensus decisions may be reached for the tumour and the normal samples, creating false positives when inferring somatic mutations.

The quality scores for the nucleotides are used extensively in the algorithms for identifying mutations and in the subsequent filters. However, the quality scores contain systematic errors associated with, for example, the sequencing machine cycle and the sequence context. A process called base recalibration has been shown to effectively reduce the bias in quality scores (DePristo et al., 2011). In Paper I-III, we applied base recalibration considering the sequence context, sequencing cycle, original base quality score and read group ID (Van der Auwera et al., 2013). The read group ID gathers reads from the same sample preparation and machine lane.

3.2 Identification of candidate somatic mutations

When trying to identify mutations, a decision has to be made whether discrepancies from the reference sequence in the reads covering a specific position reflect a true mutation or are a result of noise in the data. The statistical method used to identify somatic mutations must be sensitive to detect low-frequency mutations and mutations in regions with low sequence coverage. Simultaneously, high specificity is important due to the high dimension of the data when a large number of genomic positions are considered. Additionally, a mutation found in the tumour must be classified as somatic or germline, for which the normal sample is utilised. Typically, a statistical model is used to identify candidate somatic mutations, followed by filtering the candidate list to further remove false positives. The first step is described in this section, while the filtering part is the topic of section 3.3. First, a short overview of different methods for the identification of somatic mutations is given, followed by a description of the methods used in Paper I-III.

One method that has been used to identify candidate somatic mutations, especially in early studies, is a simple comparison between mutation lists from

tumour and normal samples (Pleasance et al., 2010). One starts with using a method for identifying germline mutations, such as, for example, the Unified Genotyper (DePristo et al., 2011), on the normal and tumour samples separately. Then, the list of mutations in the normal sample is subtracted from the list of mutations in the tumour sample. One major disadvantage of this method is that low-frequency mutations in the tumour are at risk of being missed, since the statistical model incorrectly make the assumption of heterozygous or homozygous (VAF 50% or 100%) mutations. Furthermore, all germline mutations that are missed in the normal sample but detected in the tumour will present as false positive somatic mutations.

To improve performance, a number of dedicated statistical methods for identifying somatic mutations have been developed. An overview of the available methods together with their underlying models is provided in Xu (2018). The methods can be divided into different categories, based on the assumptions and models they employ. First, there are methods that still assume diploidy in both tumour and normal samples, but instead of considering each sample separately, they perform a joint genotype analysis. SomaticSniper is an example of such a method, where a somatic score is calculated based on the probability of the tumour and normal sample having the same genotype (Larson et al., 2012). The posterior probability for each combination of genotype in tumour and normal samples, given the data, is calculated according to Bayes' rule. The prior probability of a specific combination is assumed to depend on the expected rate of heterozygous mutations in the population and the typical rate of somatic mutations. The genotype likelihood is calculated by assuming that each observation of a read is an independent Bernoulli trial with a success probability that is dependent on the genotype and the probability of an error (taken from the base quality score).

A second category of methods instead models allele frequencies and allows them to vary in a continuous range for the tumour sample. MuTect and Strelka are two examples of such models, which we used in Paper I-III and are described more in depth below (Cibulskis et al., 2013; Saunders et al., 2012). In a comparison of methods from these first categories, the ones that allow for a range of allele frequencies instead of assuming heterozygous/homozygous mutations are shown to have much higher sensitivities for detecting low-frequency mutations (Xu et al., 2014).

Furthermore, the method Varscan2, which was also used in Paper I-III and described below, is a representative from the category of methods with heuristic approaches that rely on thresholds on, for example, how many reads should support the variant allele (Koboldt et al., 2012). Methods using machine learning constitute a fourth category, and an example is SNooPer, which trains a random forest classifier to divide candidate mutations between true somatic variants and false positives (Spinella et al., 2016). A drawback is that a training set with

known true variants and errors, where the samples are prepared and sequenced in a similar way, needs to be available. Finally, methods based on haplotypes instead of individual positions in the genome have been developed (Sengupta et al. (2016), MuTect2: https://software.broadinstitute.org/gatk/documentation). In MuTect2, candidate haplotypes in regions with variation are defined via a local de-novo assembly and the reads are then aligned to the haplotypes. This process provides better accuracy in regions with several variants close to each other and facilitates calling of indels.

Several articles comparing methods for the identification of somatic mutations have been published, for example Xu et al. (2014) and Cai et al. (2016). An important remark from Cai et al. (2016) is that tuning of the model parameters and customising subsequent filters included in the methods often have a large effect on the quality and reliability of the output. It is thus recommended to acquire knowledge about the applied methods in order to adapt parameters and filters to the current application and data set.

To identify candidate SNVs in Paper I-III, we used the method MuTect (Cibulskis et al., 2013). As said above, it allows for a continuous range of possible frequencies for the sought somatic mutations in its statistical model. To detect genomic positions with a mutation in the tumour sample, MuTect applies a Bayesian classifier. Two alternative models are considered for each position harbouring variant alleles in the data, one denoted M_f^m assuming that a variant allele m with frequency f is present in the sample, and one denoted M_0 assuming that no variant alleles truly exist in the sample. The likelihood of each model is calculated based on the sequence data, taking the read nucleotides and their quality scores into account. For details on the calculation of the likelihoods, see Online Methods in Cibulskis et al. (2013). The ratio of the likelihoods times the prior probability for each model is calculated and compared to a decision threshold $\log_{10} \delta_T$:

$$\log_{10} \frac{L(M_f^m)P(m,f)}{L(M_0)(1-P(m,f))} \geq \log_{10} \delta_T.$$

The choice of δ_T determines how much more confidence that is required in the model with versus without a mutation, to declare that the position harbours a candidate mutation. By assuming a constant prior probability $P(m, f)$, the equation can be rearranged to

$$\log_{10} \frac{L(M_f^m)}{L(M_0)} \geq \theta_T,$$

where θ_T is a constant, depending on δ_T and $P(m, f)$, which can be tuned to achieve different sensitivities. When the performance of MuTect was evaluated by Cibulskis et al. (2013), a δ_T of 2 and a prior probability for a somatic mutation of 3×10^{-6} were chosen, yielding a threshold of $\theta_T = 6.3$. In Paper I-III,

we instead chose to set $\theta_T = 8$, representing both a lower prior probability for somatic mutations in the studied tumour types and a higher ratio of the likelihood needed to call a mutation. For each position with a candidate mutation in the tumour, a similar method is used for the normal data to classify the mutation as somatic or germline. The mutation frequency in the model with a germline mutation is assumed to be 0.5 (assuming heterozygosity). To assure that there is convincing evidence for *not* having a germline mutation at the position, a ten times higher likelihood for the model without a mutation is required to classify a candidate mutation as somatic. In addition, a filter for the maximum number, or proportion, of variant alleles that are allowed to be observed in the normal sample is added. In Paper I-III, we choose to reject a candidate somatic mutation if three or more variant alleles, or a proportion above 8%, were observed in the normal sample.

To identify candidate indels in Paper I-III, we used a combination of two methods, Varscan2 and Strelka (Koboldt et al., 2012; Saunders et al., 2012). In Varscan2, all positions in the normal and tumour samples are first inspected separately to assess whether a larger proportion of variant alleles is presnt in the data than a user-defined threshold. In Paper I-III, we set the threshold to 0.05. For positions where the threshold is exceeded in the tumour but not in the normal sample, Fisher's exact test is used to test for evidence of a significant difference in allele frequency between tumour and normal samples. In Strelka, a Bayesian approach is instead used. Briefly, the VAF in the normal sample is modelled as a mixture of heterozygous/homozygous genotypes and noise. The VAF in the tumour sample is modelled as a mixture of the normal sample and additional somatic variation. Thereby, a continuous range of possible frequencies for the sought somatic mutations are permitted, and base qualities are taken into account. For full details regarding the statistical model used in Strelka, see Saunders et al. (2012).

Finally, it is worth noting that before applying the statistical models described in this section, all methods have their own prior filtering regarding which positions that have enough coverage data to be evaluated and which reads that are of sufficient quality to be used. For each method utilised in Paper I-III, we used the default settings (Cibulskis et al. (2013), Saunders et al. (2012), http://varscan.sourceforge.net/).

3.3 Filtering of candidate somatic mutations

The methods used to identify candidate somatic mutations operates on data from one position at a time, assuming that the sequencing errors are random and independent, and further that all reads are aligned correctly. These assumptions are in general not met. For example, reads can be aligned at the wrong place or with wrong decisions concerning where to incorporate mis-

matches/indels, and sequence errors tend to accumulate for certain preceding sequence patterns. Thus, the whole error structure of the data is complex and not fully captured by the models, and the list of candidate somatic mutations often contains a high rate of false positives. The statistical models described above are therefore in general complemented with different approaches to filter the list of candidate somatic mutations. For example, MuTect has multiple numbers of implemented filters that we applied to the lists of candidate somatic mutations in Paper I-III (Cibulskis et al., 2013). Important steps of the filtering include removing mutations at positions with proximal gaps, i.e., where the aligned reads spanning the position harbour surrounding indels, and mutations where the variant alleles mainly sit at the start or end of reads. Furthermore, positions where the mapping quality scores are low for the reads supporting the mutation or indicate that many of the reads could have been placed equally well at another region, are also excluded. Another example is the removal of mutations with strand bias, i.e., where mismatches are seen mainly in one read direction and thus can be assumed to be dependent on the sequence context.

However, the filters added to each method do not cover all systematic errors that can occur. The paired design used when calling somatic mutations means that data from each tumour sample is compared to data from the normal sample in the same patient. The aim is primarily to exclude germline mutations, but technical artefacts present in both both tumour and normal data are also captured. To remove false positives due to rare but systematic position-specific errors, not only the paired normal sample but all the normal samples can be utilised. In Paper I-III, we used an approach where we screened all the normal samples at all positions where candidate somatic SNVs were identified. If two or more samples failed to meet the normal criteria (at most 2 reads or 8% of the reads harbouring the variant allele) at a specific position, the corresponding candidate somatic mutation was excluded. During analysis of the data in Paper I-III, we also noticed that variant alleles were sometimes identified recurrently at certain positions only for samples sequenced under the same conditions. That is, the position-specific sequencing errors correlated to the type of sequencing machine (e.g., HiScanSQ or NextSeq) and its settings, including the chemistry version. An important aspect of the study design is therefore to run paired samples together or, at least, use the same experimental set-up. Furthermore, to fully utilise the screening of normal samples, it is important to have access to as many samples as possible that are sequenced under similar conditions.

We have now arrived at a list of somatic mutations that are evaluated and filtered from a technical perspective. One remaining question concerns which of the somatic mutations that influence the disease in a crucial way and which that are merely passenger mutations. It is noteworthy that this is an important part of the aims in Paper I, while in Paper II and III, the focus is instead on the

identification of patient-specific genetic markers that are present in as many tumour cells as possible.

A first step towards elucidating the importance of mutations is to annotate them with respect to gene name, location in functional elements, if any amino acid substitution occurs and the (germline) population frequency of the mutation. In Paper I-III, we used the tool ANNOVAR to annotate the list of somatic mutations (Wang et al., 2010). Mutations with a population frequency greater than 1% were excluded from the lists. In Paper II and III, these common mutations were removed due to a higher likelihood of being missed germline mutations, which would make them and thus unsuitable as tumour cell markers. In Paper I, an additional reason for removing mutations that are common in the population was that it is unlikely that such mutations are the cause of the rare cancer disease studied herein. For mutations located in protein-coding regions, only mutations resulting in a change in the amino acid chain of the encoded protein (nonsynonymous mutations) were retained in Paper I. A change in the amino acid chain is a prerequisite for altering the function of a protein, but different changes affects the protein structure and function to different extents. As a further guidance for the functional consequences of the somatic mutations found in Paper I, we also annotated the mutations with the scores from five different functional prediction algorithms (Liu et al., 2013).

A strong criterion for the influence on the disease is whether a gene is mutated recurrently, i.e., has somatic mutations in several patients. However, when analysing a large collection of samples or tumours with a high mutation rate, recurrent mutations in a gene can occur solely by chance, especially for large proteins. There are statistical methods to test the hypothesis that a gene exhibits more mutations than expected according to the background mutation rate (Raphael et al., 2014). In Paper I, such tests on the gene level were not applicable due to the low somatic mutation rate in combination with a heterogeneous disease and rather few samples. Instead, we high-lighted all the genes that harboured recurrent mutations, with the exception of genes that were previously suggested to often represent false positives in cancer studies due to a large size or high mutation frequency (Lawrence et al., 2013).

Chapter 4

Sensitive detection of mutations using targeted deep sequencing

In this chapter, we focus on methods for quantifying low-frequency somatic mutations using sequencing data. The first section provides an introduction, with examples of applications, associated challenges and some of the key methods that have been proposed. The remaining two sections consider the methods that were applied and developed in Paper III. The application is introduced in section 4.2, and the methods used to reduce the level of noise in the data is described. In section 4.3, a novel statistical model with a position-specific error correction is presented and discussed.

4.1 Introduction to targeted deep sequencing

As described in Chapter 3, there is in general an interest in detecting somatic mutations in exome sequencing data at any frequency, including both mutations common to all cancer cells, present at frequencies near 50%, and mutations that are only present in subclones and hence show a lower variant allele frequency (VAF) than 50%. However, exome sequencing and the methods described in Chapter 3 have a lower limit with respect to the possible mutation frequency for detection with sufficient confidence. For example, in Paper I, we excluded all candidate somatic mutations with a VAF below 5% to avoid large amounts of false positives. This is both due to the available sequencing depth of the exome sequencing data, where the coverage typically is around 100, and the relatively high levels of noise. In many applications, however, there is a need

to detect mutations at lower frequencies. One example, that was already mentioned in Chapter 3, is the analysis of biopsies from solid tumours that are heterogeneous and/or include a large number of normal cells. In these cases, the fraction of cells with a specific somatic mutations can be low (Shin et al., 2017). Additionally, if the interest of a study focuses on clonal evolution, the low frequency mutations in small subclones are important (Gerstung et al., 2012).

Another example is the emerging area of analysing cell-free DNA (cfDNA) in blood samples (often referred to as the analysis of liquid biopsies). It is known that the blood plasma contains cfDNA and that the cfDNA is released from different parts of the body, including DNA from the foetus during pregnancy and DNA from tumours in cancer cases (Lo et al., 1997). Analysis of the cfDNA during pregnancy can be used in prenatal testing, with the advantage of analysis of the foetal DNA without invasive procedures such as amniocentesis (Breveglieri et al., 2019). Future possible clinical applications for analysis of circulating tumour DNA (ctDNA) includes monitoring of the disease burden, prognostic determination, selection of treatment and monitoring for relapse (Fiala and Diamandis, 2018). For example, the amount of ctDNA correlates with the tumour size, tumour stage and metastatic burden. Furthermore, knowledge concerning specific mutations in ctDNA can guide the treatment selection through knowledge about the potential drug resistance or drug targets, as well as the prognostic indication. Taking a liquid biopsy instead of images or invasive samples avoids the risks associated with radiation exposure, may provide a more holistic picture of the tumour and allows more frequent patient monitoring. However, an important feature and challenge in the analysis of cfDNA is that the proportion of DNA of interest (foetal DNA or ctDNA) is often very low, which implicates the need for methods with a level of detection down to, and even below, a VAF of 0.01% (Fox et al., 2014).

A related example with similar demands and applications to those for ctDNA, is the analysis of minimal/measurable residual disease (MRD) in leukaemia. In MRD analysis, liquid samples from bone marrow or blood are collected, but instead of the plasma, the leukocytes are analysed for the presence of low amounts of leukaemic cells. By utilising leukaemia-specific genetic aberrations, the number of leukaemic cells can be inferred from the sequencing data. The consensus requirement for MRD methods in acute myeloid leukaemia (AML), the disease in focus in Paper II and III, is that they should be able to detect leukaemic cells down to 0.1% (Schuurhuis et al., 2018), but even lower levels are desirable.

The error profile

To meet the resolution requirements for the described applications, there is a need for very sensitive methods. To be able to use sequencing data, the

depth must be increased, which can be achieved by targeting fewer and shorter genomic regions than in exome sequencing, but instead sequence those to a depth in the order of 100,000 or even higher. However, error rates up to 1%, which have been reported for Illumina sequencing, the most commonly used sequencing technique, still impose a challenge (Schirmer et al., 2016).

In targeted deep sequencing, a small part of the genome containing the position(s) of interest is first singled out and amplified by polymerase chain reaction (PCR). In this process, the genomic material from the cells in the sample is extracted and then targeted by PCR primers specifically designed to only bind to the genomic region of interest. Starting at the attached PCR primer, a copy of the desired part is built by elongating the new chain with complementary nucleotides. The resulting DNA fragments are then further amplified in additional PCR cycles. The DNA fragments are also prepared for the subsequent sequencing through incorporation of, for example, sequencing adapters and the index. The amplified fragments are then sequenced to a high depth, typically generating $10^5 - 10^6$ reads.

As described in Chapter 1, errors are introduced both in the sample preparation and during the actual sequencing (Olson et al., 2015). The use of PCR in the sample preparation means that indicates nucleotides may be incorporated. This create errors that are already present before the DNA fragments enter the sequencing machine and hence cannot be corrected for using the base quality scores that are assigned to each base during sequencing. In contrast, the well-known property that sequencing errors occur more often at the ends of reads is attributed to the accumulation of phasing and pre-phasing in the sequencing-by-synthesis method used in Illumina sequencing, and can thus be reflected in the base quality scores (Schirmer et al., 2015). Other examples of error causes in Illumina sequencing are the incorporation of incorrect nucleotides during the sequencing-by-synthesis process and incorrect base-calling when interpreting the fluorescent signals.

The errors introduced during sample preparation and sequencing come in different forms. Some errors appear independently of the underlying sequence. There are, however, also errors that are highly dependent on the nucleotide context and, thus, are much more common in certain genomic positions. For example, it has been shown that an elevated error rate is associated with the preceding motif "GG" (Schirmer et al., 2016). There are discrepant results regarding how well the assigned base quality scores actually characterise the errors introduced during sequencing (Kozich et al., 2013; Schirmer et al., 2015). Nevertheless, it can be concluded that bioinformatical approaches to remove or correct error-prone reads or parts of reads based on quality scores help in reducing the overall error rates. Schirmer et al. (2016) showed that for their data, produced using a variety of sequencing platforms and sample preparation methods, on average 69% of the errors could be removed by methods utilising

quality scores. However, the elevated error rates associated with specific motifs (i.e., nucleotide patterns) persisted. It was further shown that the systematic bias in error rate differed between sample preparation methods and sequencing platforms. Thus, the error rates in the sequencing data typically vary systematically between different genomic positions, and the patterns are specific to each experimental setup. The systematic biases in error rates will hereafter be denoted as position-specific errors.

Another source of errors in targeted deep sequencing originates from the practice of multiplexing, i.e., to sequence several different samples in the same sequencing run. To distinguish between the samples, a specific index is given to each sample by including the same short piece of DNA in all fragments originating from that sample. Mis-assignments of reads due to for example errors in the sequencing of the index or in the cluster separation lead to carry-over between samples analysed in the same run and potentially elevated error rates (Bartram et al., 2016).

Unique molecular identifiers

Different experimental techniques have been developed to reduce the error rates or provide means of correcting for errors, among which one strategy is the use of unique molecular identifiers (UMIs) (Chaudhary and Wesemann, 2018). In targeted deep sequencing, each DNA fragment is amplified during the sample preparation and, due to the large depth, several of these copies are sequenced, resulting in multiple reads originating from the same DNA fragment. In exome sequencing, reading one of the original DNA fragments multiple times is not desirable, as discussed in section 3.1, and copies are marked as duplicated reads and excluded. In deep sequencing, sequenced copies can instead be used to correct for errors that are introduced during the sequencing-by-synthesis procedure or even late in the PCR amplification during the sample preparation, hence only affecting some of the copies.

A requirement for error correction is that it can be inferred which reads originate from the same original DNA fragment. This is something that can be solved by attaching a UMI to each DNA fragment before amplification. Then, a consensus read can be inferred from all reads carrying the same UMI, ignoring discrepancies that only occur in some of the reads. Although the technique thus has the ability to reduce the error levels, a number of factors can impede an accurate analysis. For example, to obtain a representative result, tagging of the original fragments need to be uniform and efficient, and the tags should not give rise to bias in the PCR amplification. Both factors have been reported to be challenging (Kou et al., 2016). Additionally, the UMIs themselves can be affected by PCR errors, making it more difficult to determine which reads should be grouped together.

Identification of low frequency mutations

To identify low frequency mutations in deep sequencing data, the error structure must be taken into account. The existence of position-specific errors has been considered in various variant calling methods. One strategy used for exome sequencing data and somatic mutations was described in section 3.3. There, a set of other samples, sequenced under similar conditions but believed not to contain a mutation at the position of interest, were used to reveal positions with elevated error rates and hence point out potential artefacts in the list of candidate somatic mutations. For targeted deep sequencing, it is possible to infer position-specific error rates from a single control sample since the larger depth provides a higher resolution. One example where this is utilised is the somatic variant calling tool deepSNV.

In deepSNV, a tumour sample is compared with a control sample, and each considered position is tested for differences in the rate of variant nucleotides (Gerstung et al., 2012). The primary aim of the analysis is thus to determine whether somatic mutations are present. For each position i, the counts of the observed nucleotides in the control sample are modelled by a beta-binomial distribution. The beta distribution is parameterised by the mean error rate for the position (q_i) and an overdispersion parameter. The overdispersion parameter is assumed to be the same for all positions in the sequenced fragments and included to allow for extra variability in addition to the binomial one. Similarly, the counts in the tumour sample are modelled by a beta-binomial distribution, but the mean parameter in the beta distribution now includes both the mean error rate and the frequency of a potential mutation ($p_i = q_i + f$). With a likelihood ratio test statistic the alternative hypothesis that a mutation is present ($p_i > q_i$) is tested against the null hypothesis of no mutation ($p_i = q_i$). The overdispersion parameter in the likelihoods is estimated from two samples (control and tumour), assuming that the null hypothesis is true for all positions. The test is applied to each strand separately, making it possible to consider different error rates depending on the sequencing direction.

Specific statistical methods for data generated by experimental techniques using UMIs have been developed. See for example Xu (2018) for an overview of such methods.

4.2 Using sequencing to quantify MRD levels in leukaemia

In MRD analysis, the aim is to detect and quantify residual leukaemic cells in bone marrow or blood samples during or after the treatment, in order to monitor the treatment effect or detect relapses at an early stage (Ravandi et al., 2018). The task of distinguishing and quantifying the leukaemic cells can be

solved in different ways. Examples include identification with flow cytometry, which is based on the patterns of expressed proteins at the cell membrane, or quantifying fusion gene transcripts in cases where such an aberration is present (Ommen, 2016). In Paper II and III, a personalised method for MRD analysis in acute myeloid leukaemia (AML) using sequencing data and leukaemia-specific mutations was developed and evaluated. The idea is to define each patients profile of somatic mutations at diagnosis and determine which of the mutations are likely to be present in all leukaemic cells. Then, these mutations are proposed for utilisation as patient-specific markers in an MRD analysis using targeted deep sequencing. However, as described above, for this to be a sensitive and accurate method, the noise in the deep sequencing data must remain low. Therefore, specific means were taken to reduce the overall error levels.

Errors arising from carry-over between samples analysed in the same run were considered. As described above, when performing deep sequencing, multiple samples are typically processed simultaneously. This is possible due to sample indexing and increases the efficiency of the method but also introduces the risk of errors in assignments and therefore noise in the data. This can be a substantial problem if samples with the same mutation in different variant allele frequency levels are sequenced together. To reduce such errors, a number of steps were taken. For example, dual unique indexing instead of single indexing was utilised and shown to reduce the carry-over by 10-fold. Additionally, no mismatches were allowed for in the identification of the index, as opposed to the default procedure where one mismatch is accepted. For more details about the means to reduce sample carry-over, see Paper III.

Furthermore, paired-end sequencing provided the possibility of reducing errors by merging overlapping reads. This process was performed with the software PEAR (Zhang et al., 2014). In the presence of an overlap at a position and differences in the two called bases, the one with the highest quality score is incorporated in the merged read. Therefore, PEAR is able to correct for sequencing errors that are reflected in the quality score. The merged reads were then quality filtered to remove reads with a low mean sequencing quality. Additionally, the merged reads were mapped to the reference genome, and only reads with a perfect match to the reference sequence in a ten-base region surrounding the analysed position were kept. Both these procedures, quality filtering and not allowing for sequencing errors in the surrounding, remove reads that are generally error prone and hence reduce the overall error level.

4.3 A statistical model for assessment of the MRD level

To take the remaining errors into account and acknowledge that error rates differs between positions, a statistical model with position-specific error correction was developed. Assume that for each chosen leukaemia-specific mutation, the variant allele is denoted by M and the underlying VAF in the patient sample is denoted by f. After sequencing and applying a number of pre-processing steps (see section 4.2), the counts of the variant allele, y_M, and the total number of allele counts, N (i.e., the sequencing depth), were recorded for each sample. Assuming a fixed N and that fragments are picked at random and independently for sequencing, the variant allele counts within a sequencing experiment for a specific sample follows a binomial distribution. Since N is large (typically between 10^5 and 10^6), a normal approximation to the binomial distribution was applied. Hence

$$Y_M \sim \text{norm}(Np_M, Np_M(1-p_M)),$$

where p_M denotes the probability of observing the variant allele. Furthermore, assume that the probability p_M follows

$$p_M = f + \epsilon,$$

where ϵ denotes the probability of observing a variant allele due to errors in the sequencing process. The error probability ϵ was estimated by sequencing a reference sample that did not contain the variant allele and using the same protocol as for the patient sample. Note that ϵ includes sequence specific errors, i.e. errors that depends on the sequence context of the chosen leukaemia-specific mutation, and hence a reference analysed at the exact same position was used in each case. The variant allele counts and the sequencing depth observed in the reference sample were denoted y_M^{ref} and N^{ref}, respectively, and hence, the error probability estimate was y_M^{ref}/N^{ref}. Similarly, the observed VAF in the patient sample, y_M/N, was used to estimate p_M. Thus, an estimator $\hat{f}$ for the underlying VAF in the patient sample was defined as

$$\hat{f} = \frac{Y_M}{N} - \frac{Y_M^{ref}}{N^{ref}}.$$

The estimate of f is also referred to as the error corrected variant allele frequency (VAF^{EC}).

The variance of $\hat{f}$ was divided into two components. The first one, denoted σ_1^2, was the sample specific variability that follows from the random sequencing of DNA fragments. It was calculated by

$$\sigma_1^2 = \frac{p_M\,(1-p_M)}{N} + \frac{\epsilon\,(1-\epsilon)}{N^{ref}}.$$

The second component captures variability in $\hat{f}$ between samples that were not yet accounted for, arising from factors such as discrepancies in sample handling, quality and preparation as well as differences in error rates between sequencing runs. It was modelled as an additive variance component, denoted σ_2^2, and it was assumed to be common across samples with the same underlying VAF f and run under the same conditions.

Thus, the distribution of $\hat{f}$ was modelled as

$$\hat{f} \sim \text{norm}(f, \sigma_1^2 + \sigma_2^2),$$

where f and σ_1^2 are sample specific, while σ_2^2 is common for all samples.

For estimation of the sample specific parameters in sample i, let $y_{M,i}$ denote the observed number of variant alleles and N_i the sequencing depth. Let the corresponding values for a reference sample, with the same leukaemia-specific mutation as in sample i, be denoted $y_{M,i}^{ref}$ and N_i^{ref}. The estimated value of the underlying VAF f_i in sample i (the estimate also referred to as VAF^{EC}) was calculated as

$$\hat{f}_i = \text{VAF}_i^{\text{EC}} = \frac{y_{M,i}}{N_i} - \frac{y_{M,i}^{ref}}{N_i^{ref}}.$$

The first variance component, $\sigma_{1,i}^2$, is unique for each sample and depends on the underlying VAF, the error probability and the sequencing depth. For sample i, $\sigma_{1,i}^2$ was estimated as

$$\hat{\sigma}_{1,i}^2 = \frac{y_{M,i}}{N_i}\left(1 - \frac{y_{M,i}}{N_i}\right)\frac{1}{N_i} + \frac{y_{M,i}^{ref}}{N_i^{ref}}\left(1 - \frac{y_{M,i}^{ref}}{N_i^{ref}}\right)\frac{1}{N_i^{ref}}.$$

The second variance component, corresponding to the between sample variability, σ_2^2, can be estimated from a set of n samples that are run under similar conditions and with the same underlying VAF f. Estimation of σ_2^2 was performed using maximum likelihood, which gives the following equation

$$\sum_{i=1}^{n} \frac{1}{(\hat{\sigma}_{1,i}^2 + \hat{\sigma}_2^2)} - \sum_{i=1}^{n} \frac{(\hat{f}_i - f)^2}{(\hat{\sigma}_{1,i}^2 + \hat{\sigma}_2^2)^2} = 0.$$

The estimate $\hat{\sigma}_2^2$ was calculated by numerical maximisation. The value for f was either estimated from the mean of $\hat{f}_i$:s or set to a known value.

In the evaluation of the proposed MRD analysis method utilising the model described above, the coefficient of variation and limit of detection was calculated by estimating the model parameters under specific experimental settings. See the Statistical Analysis paragraph in the Materials and Methods section in Paper III for a description of how this was performed.

Here, we proposed a method to determine the frequency of a leukaemia-specific mutation through an estimator based on the difference in observed VAF between the patient sample and a reference sample. The reference sample was used to estimate the error level. By having one reference sample per sought mutation and considering the exact same genomic position as for the mutation, position-specific errors were taken into account. The counts of the variant allele within a sequencing experiment were assumed to follow a binomial distribution and, hence, to vary depending on the probability of observing the variant allele, which includes the position-specific error rate, and the sequencing depth. This was taken into consideration in a first variance component, σ_1^2. As opposed to the deepSNV method (Gerstung et al., 2012) described in section 4.1, a normal approximation was made to the binomial distribution when modelling the observed counts. Additionally, instead of incorporating between-sample variability through a beta-binomial distribution we chose to add a second fixed component, σ_2^2, to the variance. Thus, the second variance component was assumed to be independent of the sequencing depth and the position-specific error level. The latter factor is, however, confined to a low value, as mutations with an estimated error level above 0.05% were evaluated as unsuitable as MRD markers and excluded from the analysis. Estimation of the second component should be performed based on samples with the same f (i.e., frequency of the mutation in the sample) and is then valid only for this level of f.

In the proposed model, the between sample variability, captured by the second variance component, is estimated from a set of samples. This is in contrast to deepSNV, in which an overdispersion parameter is estimated from only one pair of tumour and reference samples under the somewhat questionable assumption that the overdispersion is identical along the read. Note that the choice of samples used to estimate σ_2^2 determines which between-sample variability is included. For example, one can choose to only include samples where one specific position is analysed. Alternatively, as we did when estimating the limit of detection, one can include samples where many different positions are analysed and, hence, also capture the between-marker variability. We have not taken the strand explicitly into account in the model, neither set a cut-off for the base quality score. We believe that both the requirement for perfect matches to the reference in an area surrounding the position of interest, and the merging of reads, lead to the inclusion of few bases of low quality in the analysis. This remains, however, to be confirmed. Finally, the proposed method in Paper III is formulated with the prerequisite that the position of the mutation is known. However, it should be straight-forward to generalise the method to search for new mutations, with the limitation associated with all deep sequencing that the region to target must be decided. Additionally, the PCR primers were designed to have the mutation of interest in the middle of the fragment and to have substantial overlap between the paired reads to

lower the general error rate. Thus, one should keep in mind that the sensitivity and specificity may change if searching for mutations located near the ends of the reads or using reads with less overlap between read pairs.

Chapter 5

Summary of papers

5.1 Paper I – Malignant pheochromocytomas/ paragangliomas harbour mutations in transport and cell adhesion genes

Pheochromocytoma (PCC) and paraganglioma (PGL) are rare neuroendocrine tumours located in the adrenal medulla or extra-adrenal paraganglia. Just over 10% of the patients with a primary PCC/PGL tumour develop malignant disease (Goldstein et al., 1999). The prognosis for patients with malignant disease is poor, and metastases may occur several years after removing the primary tumour. Thus, long-term surveillance of PCC/PGL patients is required. This is emphasised by the observation that although some factors that may indicate a higher risk of future malignancy are known, there is currently no reliable way to predict if a primary tumour will metastasise. Inherited mutations that predispose individuals to PCC/PGL have been characterised, but less is known about additional somatic events leading to tumour progression and malignancy.

In Paper I, the aim was to investigate somatic mutations in benign and malignant PCC/PGL tumours and to identify somatic mutations that contribute to the malignant transformation. Exome-sequencing of paired samples (normal–tumour) from four patients with benign and five patients with malignant tumours was performed. Two biological replicates were collected from each tumour, one from fresh-frozen (SF) and one from formalin-fixed paraffin embedded (FFPE) material. In exome sequencing, DNA fragments from the protein-coding part of the genome are sequenced. The resulting reads are then aligned to the human reference genome to identify its original location. Both mutations and sequencing errors will be displayed as mismatches between the reads and the reference genome. The statistical challenge is then to identify

which genomic positions truly have mutations, in contrast to those where the discrepancies only represent noise. An additional difficulty in searching for somatic mutations in contrast to inherited ones is that the variant allele frequency might be low, due to a potential heterogeneity of the tumour or contamination from normal cells. The first step in the bioinformatical and statistical analysis is to pre-process the data, with the aim to correct, or at least compensate, for errors in the data. The raw sequencing data was quality-trimmed, aligned to the human reference genome, marked for duplicates, realigned patient-wise and base-recalibrated. For the subsequent identification of candidate somatic mutations, a method called MuTect was applied, which includes a Bayesian classifier that is designed to also be able to capture low-frequency mutations (Cibulskis et al., 2013). This procedure was followed by different filters to enhance specificity and account for errors not captured in MuTect's statistical model, and functional annotation. For details about the pre-processing, identification of somatic mutations, filtering and annotation, see Chapter 3.

The resulting landscape of somatic mutations included 225 unique mutations, located in 215 genes, with a median VAF of 0.27 for SF samples and 0.29 for FFPE samples. In Figure 5.1a, the VAF for each somatic mutation is shown, grouped per sample. The average mutation rate per sample was 0.54 mutations/megabase, placing the mutation rate of PCC/PGL tumours in the lower range compared with other cancer types. A significantly higher rate of mutations in malignant tumours in comparison to benign ones was observed (Figure 5.1b). Four genes had somatic mutations in more than one patient: *HRAS*, *MYCN*, *MYO5B* and *VCL*. Mutations in *HRAS* were found in benign sporadic cases, similar to the findings in previous studies examining PCC/PGL. Recurrent mutations in *MYCN*, *MYO5B* and *VCL* are, however, novel findings in PCC/PGL and were exclusively found in malignant PGL cases. Out of these three mutations, *MYCN* is a previously known oncogene. *MYO5B* and *VCL* have functions related to cell migration, an important mechanism for malignant potential in tumours. When screening publicly available PCC/PGL datasets, three additional *MYO5B* mutations were found, two in patients with malignant disease and one in a tumour displaying pathological risk factors for malignancy. Altogether, the study contributes to the search for a set of genetic markers that predicts malignancy, which could aid in treatment and surveillance decisions and benefit the outcome for patients harbouring tumours with high malignant potential.

Furthermore, the overlap between SF and FFPE samples was in general high, with an average of 58% of the mutations found in SF samples also present in corresponding FFPE samples. This result exemplifies the utility of FFPE material in exome-sequencing studies for somatic mutations. Additionally, the unique mutations identified in each sample confirm the heterogeneity of tumours and show that biological replicates contribute to a more complete picture

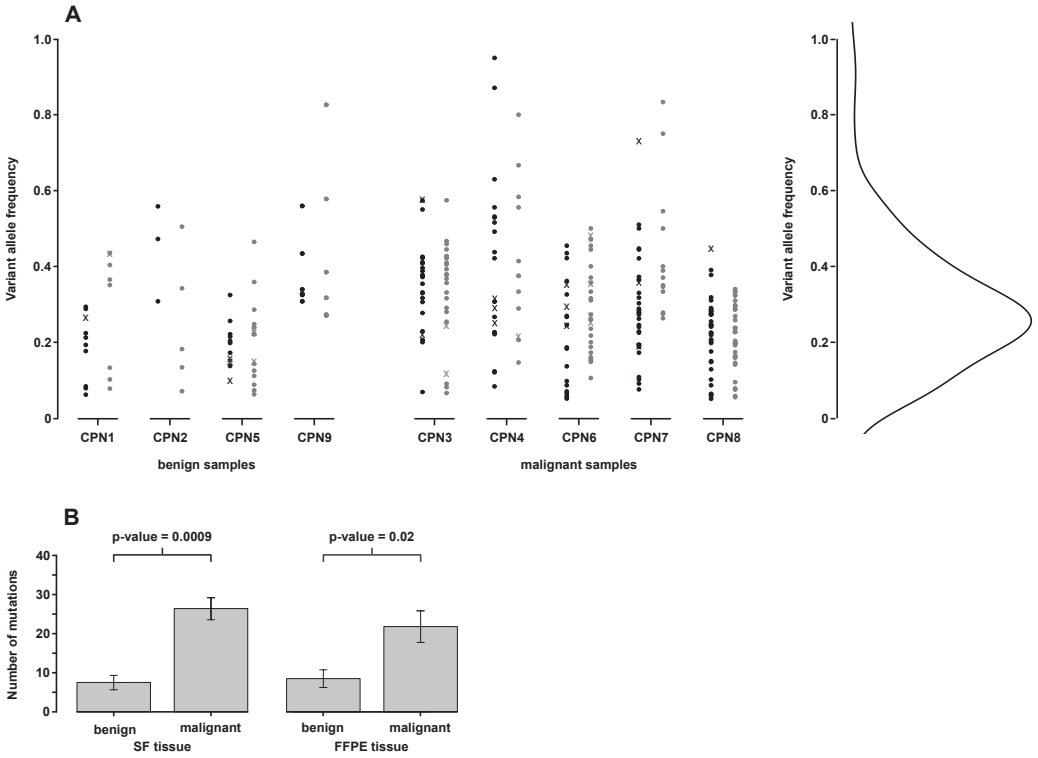

Figure 5.1: The landscape of somatic mutations in PCL/PGL. In A), the VAFs for somatic mutations in four benign cases and five malignant cases are shown. Mutations in SF samples are displayed in black, mutations in FFPE samples in grey. Dots represent substitutions, and crosses represent small insertions/deletions. The observed distribution of VAF in SF samples is shown to the right. In B), the mean number of mutations per sample is shown. There was a significant difference in the mutation rate between benign and malignant samples.

of the landscape of somatic mutations.

5.2 Paper II – Patient-tailored analysis of minimal residual disease in acute myeloid leukaemia using next generation sequencing

In leukaemia, the white blood cells are affected, and in contrast to solid tumours, the cancer cells are naturally mixed with normal cells. The most common form in adults is acute myeloid leukaemia (AML). The primary treatment is induction chemotherapy with the aim of achieving remission, i.e., no signs of the disease, followed by consolidation with chemotherapy with or without stem

cell transplantation. In the risk stratification of patients and decisions about the treatment intensity, early response to treatment is one of the most important factors. This is monitored during treatment by measuring the amount of remaining leukaemic cells, denoted the analysis of minimal/measurable residual disease (MRD). To perform MRD analysis as surveillance for patients in remission is becoming increasingly used for early detection of relapses (Ravandi et al., 2018). Today, multiparameter flow cytometry (MFC) is the most commonly used method for MRD analysis evaluating the response to treatment. It utilises the immunophenotype of the leukaemic cells, i.e., the set of expressed proteins at the cell surface. The technique is applicable in a large proportion of patients ($\sim 90\%$), but it has several disadvantages, including a potential shift in the immunophenotype, which can lead to false negative results and lowers its ability to predict relapses (Ommen, 2016). Instead, genetic aberrations in the leukaemic cells can be utilised for MRD analysis. However, the genetic heterogeneity of the disease means that there is no limited set of recurrent genetic variants that can be used for all patients (Shen et al., 2011). Therefore, patient-tailored approaches are required to utilise each patient's specific genetic changes in leukaemic cells as markers.

In the study described in Paper II, the aim was to identify leukaemia-specific mutations in patients with AML and evaluate their suitability for patient-tailored MRD analysis. To obtain the profiles of leukaemia-specific mutations in individual AML patients, leukaemic cells and normal lymphocytes from 17 patients were isolated using fluorescence activated cell sorting. The two fractions from each patient were then exome-sequenced separately and analysed in a paired design to decide which mutations existed only in the leukaemic cells. After data pre-processing, the identification of candidate leukaemia-specific mutations and filtering to remove potential false positives, see Chapter 3 for details, a total of 262 leukaemia-specific SNVs and indels were found. To avoid false negative results due to subclonality, we want candidate mutations to be present in all leukaemic cells to be classified as suitable for MRD analysis. The majority of the identified mutations had a variant allele frequency (VAF) of approximately 0.5, corresponding to being present as heterozygous mutations in all leukaemic cells (Figure 5.2a). The random selection of DNA fragments for sequencing give rise to binomial distributed variant allele counts, and observed VAFs are therefore expected to fluctuate around 0.5, even if each mutation truly is heterozygous and present in all cells. However, some mutations showed a VAF considerably lower than 0.5, which could indicate that they were only present in a subset of the leukaemic cells. A comparison of the observed VAF distribution to a simulated distribution was conducted, taking the realised sequencing depths into account and assuming heterozygosity for all mutations in the simulation (Figure 5.2b). Although the results showed an overall correspondence, there was evidence that a portion of

the observed mutations actually had a lower observed VAF than expected for heterozygous mutations. To remove those mutations unlikely to be present in all leukaemic cells, a 95% confidence interval around the VAF of each mutation was calculated based on the observed depth and utilising a normal approximation. Mutations where the interval was below 0.50 were excluded. In total, 191 leukaemia-specific mutations passed this filtering step and were thus considered candidates for MRD analysis. All patients but one had MRD candidates in their somatic mutation profile (median 11 per case, range 0-25).

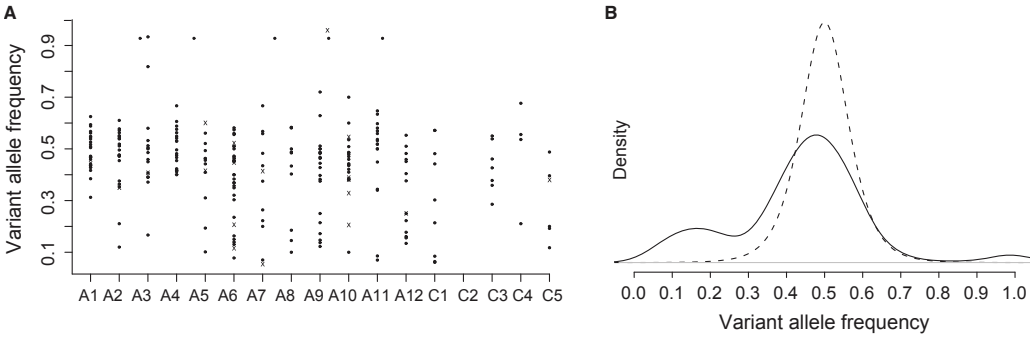

Figure 5.2: The variant allele frequencies of the identified somatic mutations. In A), the VAF for each mutation is displayed, divided by case. Dots represent substitutions and crosses represent small insertions/deletions. A denotes adult AML cases, and C indicates childhood AML cases. In B), the solid line shows the observed distribution of VAF, while the dashed line shows the simulated observed distribution when assuming all mutations to be heterozygous and present in all cells.

To detect the low frequencies of mutations that are desirable in MRD analysis, targeted deep sequencing, where a specific part of the genome is selected and sequenced to a high depth, can be utilised. The technique was used on follow-up samples from a patient with AML. Four mutations from the set of previously identified MRD candidates for the patient were analysed. The results showed that this approach for MRD analysis was more sensitive than the standard MFC method. Furthermore, when the MFC method failed to correctly capture a relapse after 10 months due to a change in immunophenotype for a majority of the leukaemic cells, all four of the mutations were detected with a high mutation load. Thus, targeted deep sequencing of the mutations, identified by exome sequencing of sorted leukaemic cells and lymphocytes, herein successfully accomplished MRD quantification. This approach could contribute to making MRD analysis possible for all patients.

5.3 Paper III – Accurate and sensitive analysis of minimal residual disease in acute myeloid leukaemia using deep sequencing of single nucleotide variations

As described in the summary for Paper II, minimal/measurable residual disease analysis is important in risk stratification and monitoring of acute myeloid leukaemia patients. In general, each patient has a number of somatic mutations in his or her leukaemic cells, which may be used as markers in MRD analysis and thus allow patient-tailored assays.

The aim of Paper III was to characterise and validate the use of targeted deep sequencing of leukaemia-specific substitution mutations, identified by exome sequencing, for MRD analysis. The basic idea of deep sequencing is that the proportion of reads having the leukaemia-specific variant allele, in contrast to reads from the same region having the reference allele, contains quantitative information about the level of leukaemic cells in the sample. The accuracy of the quantitative information is, however, challenged by errors in the data. Hence, a key aspect in using targeted deep sequencing of single nucleotide substitutions for MRD detection and quantification purposes is to keep the level of noise low to reach the consensus requirement for molecular MRD analysis of being able to detect leukaemic cells down to the 0.1% level (Schuurhuis et al., 2018). An additional aim is to develop an even more sensitive method, that is able to quantify low levels of leukaemic cells. This achievement could enable a more detailed evaluation of the results of treatment and contribute to the detection of emerging relapses as early as possible.

To reduce the overall level of noise and thereby increase the sensitivity in the assay, a number of considerations were made in the sample preparation and bioinformatical processing of the data (see section 4.2). However, it is known that sequencing errors in the MiSeq platform do not occur completely randomly. The nucleotide context for a position influences its error rate, and the three possible substitution errors for a position are not evenly distributed (Schirmer et al., 2016). To be able to take the position-specific errors into account, without making precedent general assumptions about the error structure, a reference sample for each leukaemia-specific mutation was sequenced and used for adjustment of the variant allele frequency (VAF) estimate. That is, an error corrected VAF (VAF^{EC}) was calculated as the difference in VAF between the sample of interest and a reference sample. In this way, a correction was made for both general and sequence-specific errors. A statistical model for the estimation with position-specific error correction was formulated. The variability of VAF^{EC} was divided into two components, taking both within- and between-sample variation into account. The first variance component mod-

els the sample-specific variability that follows from random sequencing of DNA fragments. The second variance component corresponds to the between-sample variability caused by factors such as discrepancies in sample handling and quality, preparation and error rates between sequencing runs. For details about the statistical model, see section 4.3.

First, the effect of the position-specific error correction was evaluated by estimating the total variability in the assay with and without the use of a reference sample. The sample standard deviation for 15 normal samples corresponding to 15 different mutations was calculated. The value without using the position-specific error correction was 0.0123%, as compared to a value of 0.0060% using VAF^{EC}. Hence, the noise level was reduced by 51%, and it can be concluded that the sensitivity of the method was substantially improved.

The precision and accuracy of the assay were then determined from DNA samples with known levels of mutations, where each set of level (0.01% and 1%) and mutation (four different mutations) was sequenced in triplicates, and the model parameters were estimated. The precision was evaluated by calculating the coefficient of variation (CV) for each set, resulting in a median value of 4.1% at VAF 1% and 13.3% at VAF 0.1%. The median relative bias at VAF 1% was 7.9%.

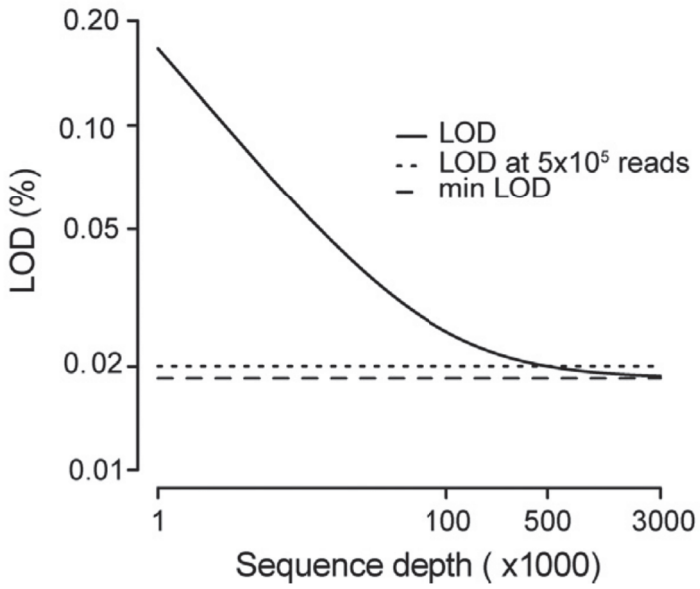

Figure 5.3: The limit of detection (LOD) as a function of sequencing depth. The LOD was estimated to a value of 0.020% at a sequencing depth of 5.0×10^5 (dotted line), utilising 15 normal samples each sequenced for a different mutation. The minimum LOD, i.e., the value that can be reached in the absence of within-sample variability, was estimated to be 0.0185% (dashed line).

The limit of detection (LOD) was determined from 15 normal samples (i.e. $f = 0$), each sequenced for a different mutation, and calculated as the absolute value of the mean VAF^{EC} + 3 standard deviations. The limit of detection depends on the sequencing depth through the within-sample variability captured in the first variance component σ_1^2, and hence, it is lower at a higher sequencing depth (Figure 5.3). At a sequencing depth of 5.0×10^5, the LOD was estimated to be 0.020%, corresponding to 1 cell in 2500 with a heterozygous mutation. Raising the sequencing depth further only lowered the LOD marginally, and the minimum value of the LOD was estimated to 0.0185%. The probability of observing a VAF^{EC} above the LOD at an underlying VAF in the patient of $f = 0.05\%$, at a sequencing depth of 5.0×10^5, was calculated to be >99.9%. That is, if a sample contains at least 0.1% leukaemic cells with a heterozygous mutation at the analysed position, the probability of detection with the assay is greater than above 99.9%. For details about the calculations of CV, LOD and probability, see the paragraph Statistical Analysis in the Materials and Methods section in Paper III.

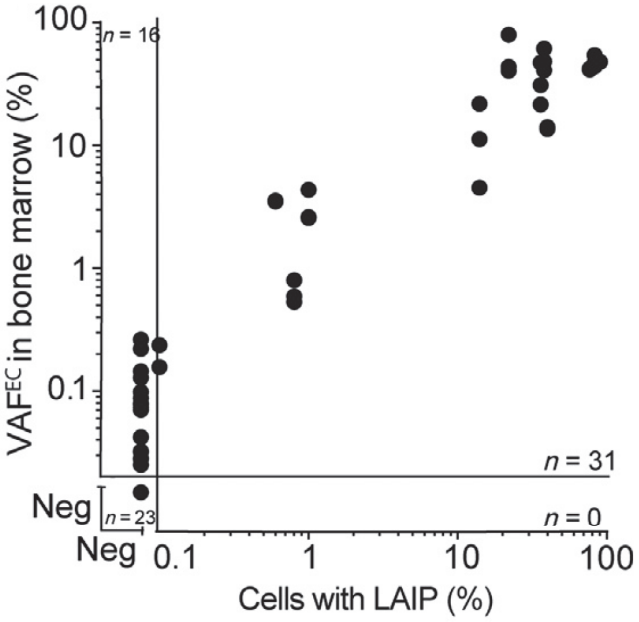

Figure 5.4: Comparison of determined MRD levels between targeted deep sequencing (y-axis) and MFC (x-axis). The lines shows the limit of detection at 0.020% for deep sequencing and 0.1% for MFC. Neg denotes a value below the limit of detection, LAIP denotes leukaemia-associated immunophenotype. Fifty-four determinations showed consistent MRD assignment (31 MRD^+ with both methods and 23 MRD^- with both methods), while 16 determinations differed (all those were MRD^+ with deep sequencing and MRD^- with MFC).

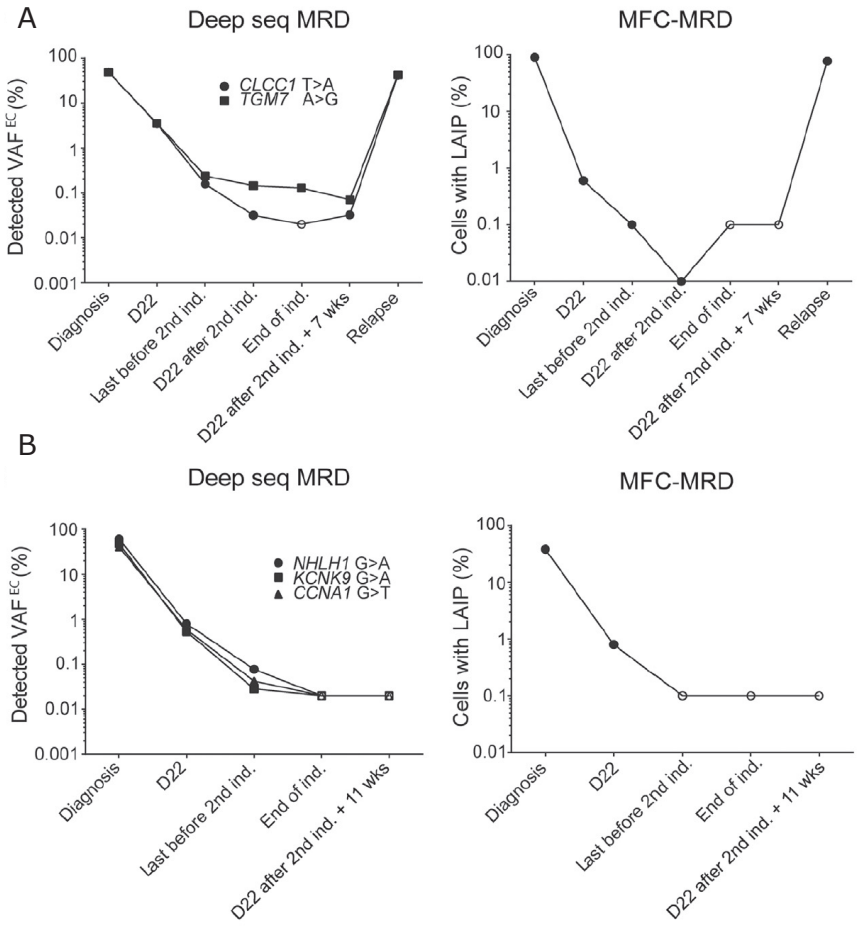

Figure 5.5: Comparison between the use of deep sequencing and MFC for MRD monitoring. In A), time series for a patient who relapsed is shown, where deep sequencing showed the MRD^+ status at the last time point before the relapse. In B), time series for at patient that remained in remission after completion of therapy is shown. D22 denotes day 22 after start of the induction course, and ind. denotes induction. Filled circles indicates measurable levels, empty circles indicates a value below the limit of detection.

The proposed method for MRD marker identification at diagnosis and quantification during treatment was tested in 6 AML cases. In total, 34 bone marrow samples were analysed for MRD levels with two or three leukaemia-specific mutations utilised per sample. Concordant MRD assignment (MRD^- if VAF^{EC}<LOD, otherwise MRD^+) for all the mutations analysed in one sample was observed in 19 of 23 samples (3 mutations analysed) and in 10 of 12 samples (2 mutations analysed). For 27 of the samples, MRD analysis using multiparameter flow cytometry (MFC) had been performed as part of the clin-

ical routine. In Figure 5.4, the levels determined with deep sequencing were compared against the levels determined with MFC. An integrated assignment of the MRD status for samples analysed with deep sequencing was defined as a sample being MRD^+ if at least one mutation showed an MRD level above the LOD and MRD^- otherwise. When comparing the MRD status based on MFC and deep sequencing, consistent results were observed in 18 out of 27 samples. The remaining 9 samples were MRD^+ with deep sequencing and MRD^- with MFC, and notably, no samples classified as MRD^+ with MFC were failed to be detected by deep sequencing (see Table 3 in Paper III). Hence, the deep sequencing assay showed concordance with MFC but had higher sensitivity. In Figure 5.5, time series for two patients with AML are shown, reporting the results for MRD analysis using deep sequencing of leukaemia-specific mutations and MFC at the same time points (6 time series in total are reported in Paper III). Note that the patient in Figure 5.5a with relapse was determined as MRD^+ with deep sequencing after the second induction while MRD^- using MFC, indicating that the higher sensitivity of the deep sequencing assay might be of prognostic value. The specificity of the deep sequencing assay was estimated to 97% by analysing 10 follow-up samples from 5 AML patients for 3 mutations that were not detected in their genomes, but instead present as leukaemia-specific mutations in another AML case.

In conclusion, patient-tailored targeted deep sequencing with correction for position-specific errors, and utilising a statistical model that takes both within- and between-sample variability into account, was shown to be able to achieve a low limit of detection, estimated herein as 0.020%. The assay thereby had superior sensitivity than ordinary MFC-MRD analysis, and it also showed high specificity. Thus, the introduction of this method in clinical care could contribute to providing virtually all AML patient with a sensitive and accurate method for MRD monitoring.

5.4 Paper IV – miRNA profiling of small intestinal neuroendocrine tumours defines novel molecular subtypes and identifies miR-375 as a biomarker of patient survival

Small intestinal neuroendocrine tumours (siNETs) are hormone-producing tumours localised in the small intestinal, with a low proliferation rate and grade, meaning that the primary tumour often grows slowly and shows resemblance to normal cells. However, the diagnosis of siNETs is often not made until metastases have developed, and the 5-year survival for patients with liver metastases is poor. The individual prognosis is, however, highly variable and difficult

to predict. Thus, there is a need to better understand the molecular mechanisms that drive progression and, hence, to identify molecular subgroups of the tumours and biomarkers for prognosis, disease progression and response to treatments. Changes in gene expression are known to be an essential component of cancer development. One form of transcriptional regulator is the microRNAs (miRNAs), which are short non-coding RNA stretches that bind to mRNA and thereby suppress translation and increase mRNA-degradation. One miRNA can regulate several different mRNAs, and they can contribute to tumour development by targeting genes involved in molecular mechanisms of cancer, e.g., cell cycle control and invasion.

In Paper IV, the aim was to determine the miRNA profile of siNETs and search for miRNA-based molecular subgroups and biomarkers. The miRNA expression was determined by one-colour microarrays in 42 tumours from 37 patients, all with a documented survival time and with well-characterised clinicopathological factors. In addition, 6 samples from normal small intestinal mucosa were analysed. Each microarray contained 866 human and 89 human viral miRNAs, measured by 2-4 different probes per miRNA. After quality control of the arrays, background correction using the normexp method was performed to remove differences in the ambient signal (Ritchie et al., 2007). Then, to enable comparison of the expression values between arrays, normalisation to remove systematic trends in the data that do not depend on the biology but rather on technical aspects was applied over all arrays using the quantile-quantile method (Bolstad et al., 2003). Spots measuring the same probe (4-8 spots per probe) were then merged on each array by taking an average of the normalised values.

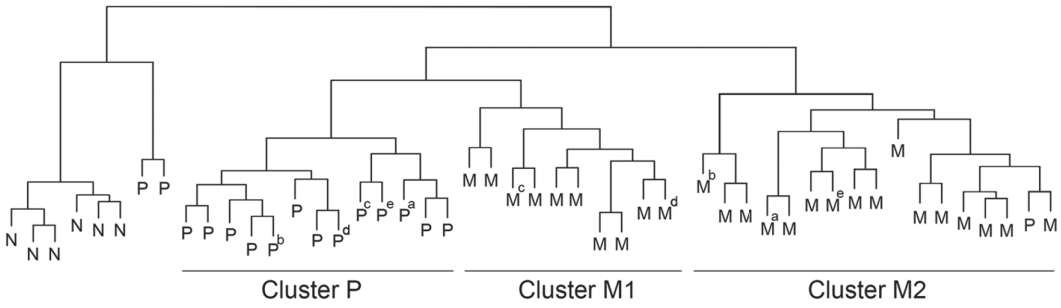

Figure 5.6: Hierarchical clustering of 42 tumour biopsies and 6 normal samples. Three major clusters of tumours are formed, one with primary tumours (Cluster P), one smaller cluster with metastases only (Cluster M1) and one larger cluster with almost exclusively metastases (Cluster M2). N denotes normal small intestinal samples, P denotes primary tumours, M denotes metastases and the letters a-e indicates tumours from the same patient (one primary tumour and one metastasis from each patient).

Clustering of the miRNA profiles of all tumours and the normal samples

was performed using unsupervised hierarchical clustering with complete linkage and the Euclidean distance as the metric. Three major clusters were observed for the tumours, one containing only primary tumours, and two clusters (denoted M1 and M2) consisting almost exclusively of metastases (Figure 5.6). Five patients contributed both primary tumours and metastases, and all paired samples were separated into different clusters. The clustering was repeated for metastases only (26 tumours), showing the same subgrouping. Test of association between clusters and the clinicopathological characteristics of the patients showed that the tumours in cluster M1 and M2 significantly differed in proliferation rate, as measured by tumour grade and Ki67 index, and chromosomal copy number (see Table 2 in Paper IV). Additionally, the patients in cluster

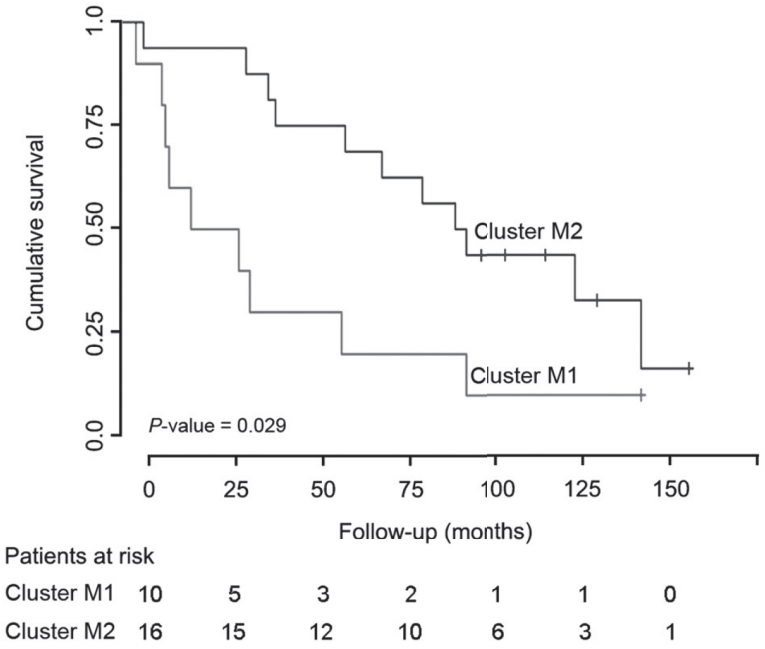

Figure 5.7: Kaplan-Meyer estimates showing the difference in survival times between patients with tumours in cluster M1 and those with tumours in cluster M2. Patients in cluster 1 had significantly shorter survival time (log-rank test, p-value 0.029).

M1 had significantly shorter overall survival (Figure 5.7). Using a Cox proportional hazards model, adjusting for age and gender and calculating the survival time from the date of surgery, the hazard ratio was estimated to 3.4 (p-value 0.018).

To identify and rank differentially expressed miRNA between two groups of samples, the moderated t-statics was used (Smyth, 2004). In principle, for each miRNA the mean difference in expression between the groups was calculated and then divided by its standard error. When estimating the standard error,

using the moderated t-statics means that information is shared between all miRNAs using a common prior distribution instead of, as for the ordinary t-statistic, only using data from the miRNA in question. The p-values in each comparison were adjusted for multiple testing using Benjamini-Hochberg false discovery rate, and tests with adjusted p-values below 0.05 were considered significant. The latter means that the expected rate of false positives among the significant miRNAs is estimated to be at most 5%. The analysis of differentially expressed miRNA between cluster M1 and M2 identified a number of miRNA previously reported to be associated with malignant behaviour of tumours. For example, miR-1246 and miR-663a, which are known to have oncogenic effects, had elevated levels in the cluster with higher proliferation rate and shorter overall survival (M1), while miR-488-3p, shown to be a tumour suppressor in gastric cancer, had reduced levels in the same cluster.

Differential expression was also examined between groups of metastases based on the proliferation rate (measured as the Ki67 index) and then based on whether the metastases had a gain of chromosome 14. Furthermore, the relationship between miRNA expression and tumour progression was investigated in a paired analysis with both primary tumours and metastases from 5 patients. For significantly regulated miRNAs and subsequent miRNA target scan followed by pathway analysis, see the Results section in Paper IV.

Comparison of miRNA expression between normal small intestinal mucosa and siNETs showed that miR-375 was upregulated in tumours. Interestingly, when searching for miRNA associated with survival time in metastases, downregulation of miR-375 was the top candidate (adjusted p-value 0.093). These findings were further investigated in situ by utilising a tissue microarray from an independent cohort of siNETs where in situ hybridisation with a miR-375 specific probe was performed. Expression of miR-375 was found in 91% of the biopsies (578/635), but only in the enteroendocrine cells located in the crypt and on the villus in normal small intestinal mucosa. Patients with higher expression of miR-375 (score 3) in liver metastases had significantly longer survival times than those with lower expression of miR-375 (score 0, 1 or 2) (Figure 5.8). When applying a Cox proportional hazards model, adjusting for age and gender, the hazard was a third for the patients with high versus low expression of miR-375 (HR: 0.32, p-value 0.026). A similar trend was observed in lymph node metastases, although it was not significant. These results are in accordance with other studies identifying miR-375 as a tumour suppressor and biomarker of prognosis. However, miR-375 has also been assigned oncogenic effects in a few cancer types, and its functional role in siNETs need to be further investigated and established.

To summarise, clinically relevant novel subgroups of metastatic siNETs were identified based on miRNA expression profiles. In the search for biomarkers in siNETs, downregulation of miR-375 showed an association with a shorter

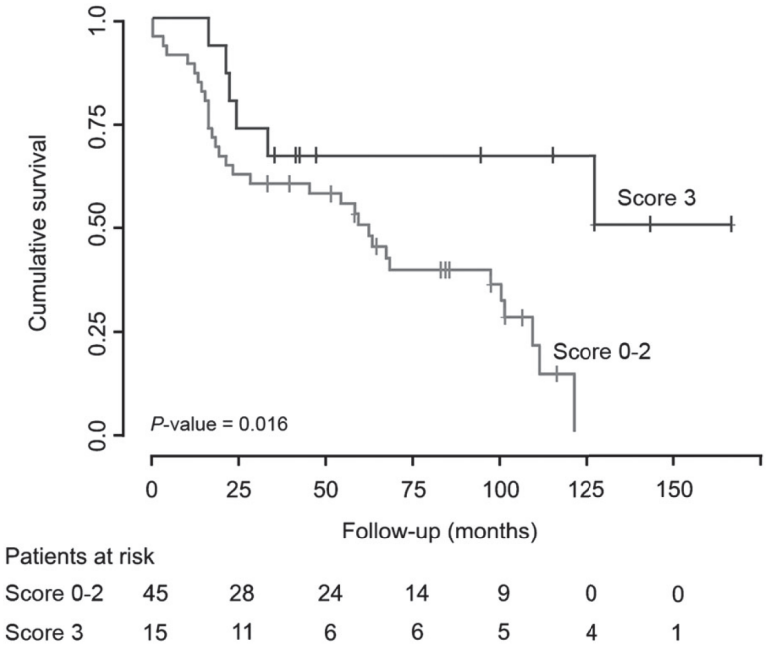

Figure 5.8: Kaplan-Meyer estimates showing the difference in survival times between patients with liver metastases with high expression of miR-375 (score 3) and low expression of miR-375 (score 0-2), as measured by in situ hybridisation on a tissue microarray. Patients with lower expression of miR-375 had a significantly shorter survival time (log-rank test, p-value 0.016).

survival time and is suggested for further evaluation as a prognostic biomarker.

5.5 Paper V - A hierarchical Bayesian model for assessing differential nucleotide composition between metagenomes

In metagenomics, microbial communities are analysed by sequencing their genetic material. Up until recently, the main interest of metagenomic studies have been the composition of species and biological functions and their variability between different experimental conditions. However, the increasing resolution in the data enables the analysis of changes in nucleotide composition, something that is crucial in order to understand the cause of phenotypic differences between microbial communities. For example, alteration of only a few nucleotides is enough to make many bacteria highly resistant to antibiotic treatment (Blair et al., 2015). The challenges in performing analysis of nucleotide differences in sequencing data include the large dimensionality of the data, with many

nucleotide sites to analyse, and the often high levels of biological and technical variability. For instance, samples from the same condition can vary considerably due to factors such as lifestyle and sex if studying the human microbiota, or temperature and available nutrients if studying environmental samples (Yatsunenko et al., 2012).

In paper V, the aim is to detect differences at nucleotide level between groups of metagenomes sampled from different experimental conditions. In particular, assuming that the average nucleotide proportions for a genomic position i in each condition respectively are described by the vectors $\tau_{i1} = [\tau_{i1}^A, \tau_{i1}^C, \tau_{i1}^G, \tau_{i1}^T]$ and $\tau_{i2} = [\tau_{i2}^A, \tau_{i2}^C, \tau_{i2}^G, \tau_{i2}^T]$, the aim is to identify the positions where τ_{i1} and τ_{i2} differ. We propose a Bayesian hierarchical model, that accounts for both within- and between-sample variability and utilises a shrinkage approach in the variance estimation.

Assume that each sample have been sequenced separately, and that vectors of counts of the four different nucleotides A, C, G and T are observed, i.e., let

$$x_{ijk} = [x_{ijk}^A, x_{ijk}^C, x_{ijk}^G, x_{ijk}^T],$$

where $i = 1, \ldots, n$ denotes the position, $j \in \{1, 2\}$ the condition and $k = 1, \ldots, m_j$ the sample. Assuming that the DNA fragments are picked randomly and independently for sequencing, the vector of counts is modeled with a multinomial distribution with parameters p_{ijk}, denoting the nucleotide composition vector (with elements summing to one), and N_{ijk}, denoting the sequencing depth. To take the biological and technical variability between samples into account, p_{ijk} is assumed to follow a Dirichlet distribution with parameters τ_{ijk}, the average nucleotide composition, and A_i, controlling the amount of extra variability compared to a standard multinomial distribution. Thus, the unconditional distribution of X_{ijk} follows a Dirichlet-multinomial distribution, where the variance is

$$\text{Var}[X_{ijk}^l | A_i, \tau_{ij}] = N_{ijk}\tau_{ij}^l(1 - \tau_{ij}^l)\left(1 + \frac{N_{ijk} - 1}{A_{ij} + 1}\right).$$

As a result, the variance has one part that is due to the random sampling of DNA fragments within a sample, and a second part that captures the between-sample variability. We apply a shrinkage approach where we let $A_i' = 1/A_i$ follow a common prior distribution, that does not depend on the position. The model was fit to data using Markov Chain Monte Carlo (MCMC) simulations with the Gibbs sampler algorithm.

In order to rank the positions according to the likeliness of a difference in genetic composition, we first let δ_i be the difference in nucleotide composition, i.e., $\delta_i = \tau_{i1} - \tau_{i2}$. A score D_i, based on measuring the distance from the posterior mean of δ_i to zero, while correcting for the uncertainty, is defined as

$$D_i = \mu_{\delta_i} \Sigma_{\delta_i}^{-} \mu_{\delta_i}^T,$$

where μ_{δ_i} and Σ_{δ_i} denotes the expected value and the covariance matrix of the posterior distribution of δ_i, respectively.

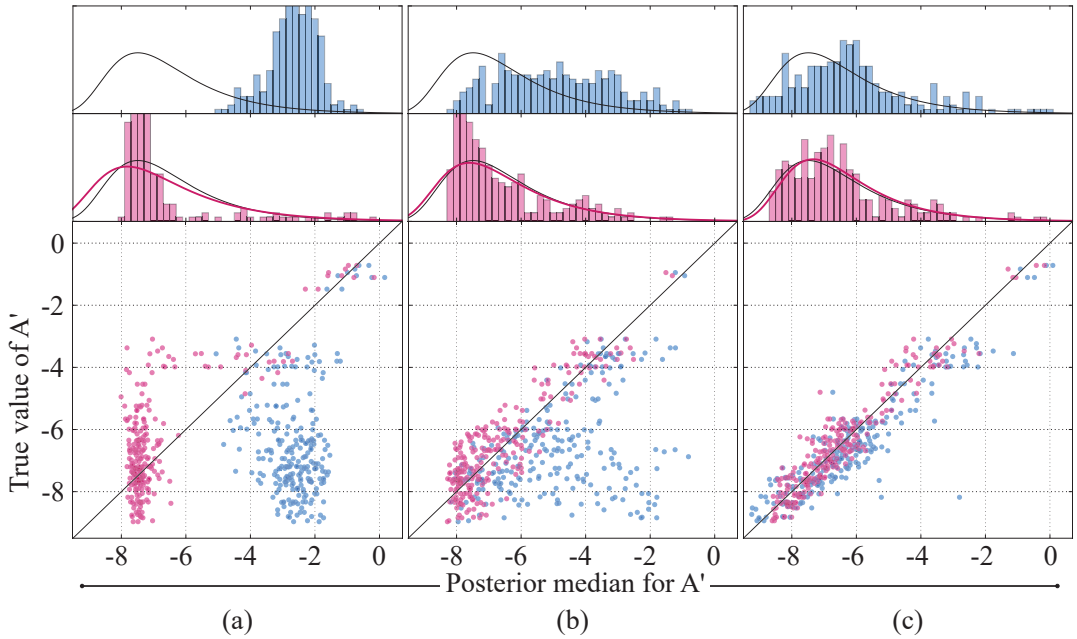

Figure 5.9: Posterior median values for the overdispersion parameter A_i' (log scale). The group size was set to 5, while the sequencing depth was varied between samples but with a mean sequencing depth that was in a) set to 100, in b) set to 1000 and in c) set to 10,000. The scatter plots show a comparison between the estimated and the true values (full model in pink, model without shrinkage in blue). The distributions of the posterior median values are shown in histograms. The red curve illustrates the fitted shrinkage distribution for A_i'. As a reference, the same curve for the full data (group size 14, mean sequencing depth 5000), from which the parameter values in the simulation was sampled, is shown in black.

The proposed model was evaluated using both simulated and real data. First, we investigated the model's ability to estimate the overdispersion parameter A_i', and the influence of the shrinkage approach. The evaluation was based on simulated data, where the the average nucleotide compositions and the overdispersion parameter were, for each position, set to values encountered in real metagenomic data. The estimated overdispersion parameter values followed the true values closely for a group size of 5 and a sequencing depth of 10,000 (Figure 5.9). When lowering the sequencing depth to 1000, the overdispersion parameter values showed larger deviance from the true values. However, with help from the common prior distribution the correlation to the true values was still evident. At the same parameter settings, the uncertainty in the estimation of the overdispersion parameters was compared between the full model and the model without shrinkage. The full model had much lower spread in

its posterior distributions for A_i' (when estimated without shrinkage, the mean posterior standard deviation for A_i' was 16 times higher). This was likewise reflected in the uncertainty in the estimation of differences in nucleotide proportions (δ_i^l), where using the shrinkage approach resulted in lower mean posterior standard deviations for all studied cases, see Table 1 in Paper V. When varying the group size while holding the sequencing depth fixed at 10,000, it could be seen that at a group size of 10 the model performed almost equally well with and without shrinkage while at a group size of 3 the overdispersion parameters totally lack information for estimation without a shrinkage approach (Figure 2 in Paper V). This shows that the proposed model, can accurately estimate the overdispersion, even if there are few samples and low sequence coverage. The use of a common prior for the overdispersion parameters reduced both the deviance from true values and the uncertainty in the estimation.

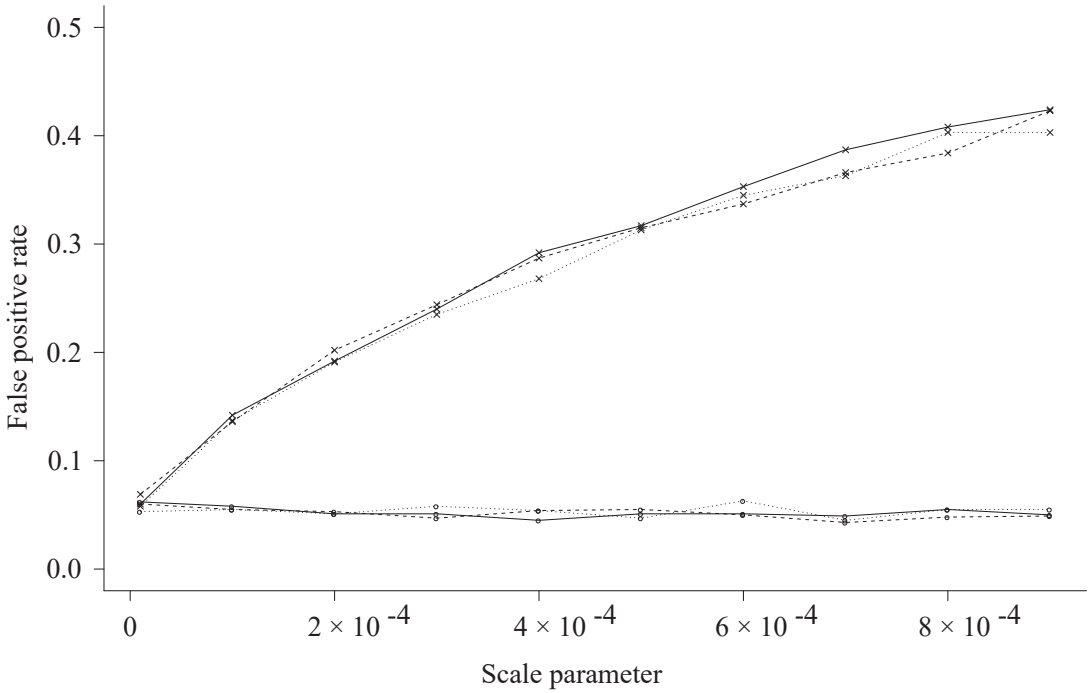

Figure 5.10: False positive rate (FPR) as a function of overdispersion, the latter represented by different values for the scale parameter in the inverse gamma distribution for the overdispersion parameter. Two different means of ranking and three different group sizes are displayed: proposed model (circles), χ^2-test (crosses), group size 3+3 (dotted line), 5+5 (dashed line) and 10+10 (solid line). The sequencing depth was varied between samples, with a mean depth set to 1000. The proposed model was able to maintain a low FPR for all levels of overdispersion. This was in contrast to the χ^2-test where the FPR increases rapidly when overdispersion was increased.

Next, the model's ability to control the false positive rate (FPR) was exam-

ined. When increasing the values of the overdispersion parameters, the FPR was shown to be kept at a consistently low level for the proposed model (Figure 5.10). On the contrary, when instead applying a χ^2-test, the FPR increased to unsatisfactory levels. Thus, standard methods, such as the χ^2-test, which does not take the overdispersion in the data into account, can result in a very high false positive rate.

The sensitivity of the model, measured as the ability to place the positions with an effect at the top of a ranking list, was evaluated using resampled data. Samples from a publicly available dataset was randomly drawn to form groups of a specific size, and effects were added to 30 out of 237 studied positions. Also, additional variability were added to some datasets, in order to evaluate the performance at different levels of overdispersion. First, the ranking performance at a sequencing depth of 1000 and a group size of 5 was examined (Figure 4 in Paper V). As expected, the performance was increased with larger effect size. Nevertheless, even for an effect size as low as 0.01, with low overdispersion all the 20 top-ranked positions had a true effect and the true discovery rate (TDR) was estimated to 0.85 at position 30 in the ranking list. With increased overdispersion, the ranking performance was reduced. At an effect size of 0.05, the estimated TDR at position 30 was 0.98, 0.84 and 0.71 for low, intermediate and high level of overdispersion, respectively. The ordinary χ^2-test showed considerably lower performance for all settings, and especially suffered at higher levels of overdispersion (Figure 4 in Paper V).

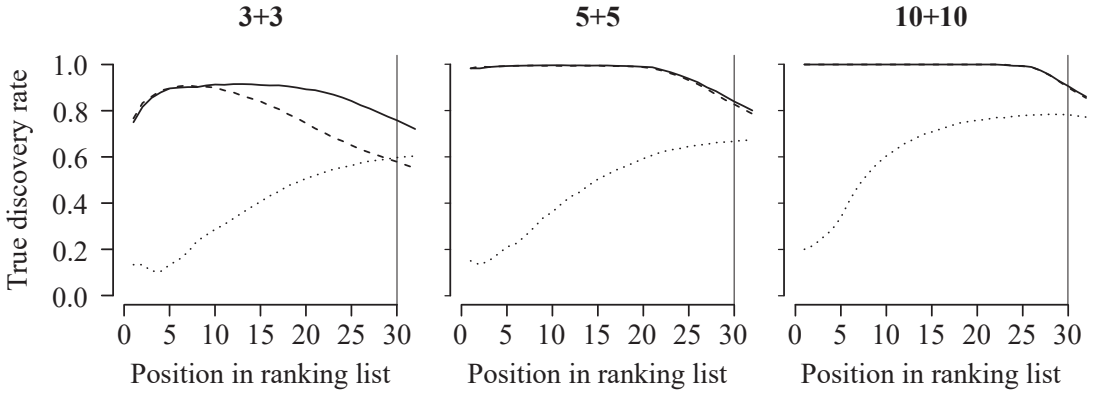

Figure 5.11: True discovery rate as a function of position in the ranking list. The results were estimated from resampled data, where the mean sequencing depth was set to 1000, the average overdispersion level to intermediate and the mean effect size to 0.05. The group size was set to 3, 5 and 10 for the plot to the left, middle and right, respectively (indicated as 3+3, 5+5 and 10+10). Three different rankings are displayed: D_i based on the full model (solid line), D_i based on the model without shrinkage (dashed line) and χ^2-test (dotted line).

The performance of the model was also evaluated for different group sizes,

showing an increased ranking ability for larger group sizes (Figure 5.11). At a group size of 3 and a sequencing depth of 1000, there was a substantial difference in ranking performance between the full model and the model without shrinkage. Lowering the sequencing depth to 100 (at a group size of 3) made the performance even more dependent on the shrinkage approach, while for a sequencing depth of 10,000 there was only a small difference in performance between the models (Figure 6 in Paper V). Together, this shows that the proposed model has a high sensitivity, even when the differences are small and the number of samples low.

Finally, the model was tested in a case study, comparing the nucleotide composition in communities of the bacteria *Escherichia coli*, before and after travelling to India (Johnning et al., 2015). Data from the gene gyrA, which is related to antibiotic resistance, was analysed. The model was able to pin-point an increase of a mutation known to cause resistance against fluoroquinolone antibiotics (see Table 2 in Paper V). This shows that travelling can promote the spread of antibiotic resistant bacteria.

In this study we show that it is possible to identify mutations also in the complex datasets that microbial communities constitute. The proposed model had a high ability to identify positions with a difference in nucleotide composition and to keep the false positive rate low, also under increased levels of between-sample variability. The shrinkage approach made it possible to draw accurate conclusions from sparse data, and reduced the uncertainty in the estimation. We conclude that tailored statistical models are important when interpreting metagenomic sequencing data and can aid in increasing our understanding of complex microbial communities.

Chapter 6

Conclusions and outlook

In this thesis, bioinformatical and statistical methods for analysis of high-dimensional data have been developed and applied in different biological and medical applications. Several types of molecular information have been utilised, including data from high-throughput DNA sequencing and large-scale gene expression profiling using microarrays. A common challenge in all applications has been to understand and properly model the variance structure in the data, and thereby be able to separate the biological signal from the noise.

First, a method for identifying somatic mutations in exome sequencing data from paired tumor–normal samples was established. By applying noise-reduction methods adapted to the study design and statistical models dedicated to detect somatic mutations, candidate nucleotide positions were identified in a sensitive way. The false positive rate was reduced by careful inspection of each candidate and custom-made filters based on the knowledge about the error structure in the data. This resulted in, for example, identification of both previously known and novel genes with recurrent somatic mutations in PCC/PGL, with potential impact on malignancy. One of the novel genes, MYO5B, has been further examined in functional studies. We could show that cell lines transfected with MYO5B genes harbouring the found somatic mutations had a significantly increased proliferation and migration rate (data not yet published). In addition, we recorded a higher mutation rate in malignant cases compared to benign cases, something that has not been reported before in PCC/PGL but has been encountered in other cancer forms. The results indicate that the methods that were applied for identification of somatic mutations have both the necessary sensitivity and specificity.

It can be noted that results regarding the rate of somatic mutations in different cancer types and subgroups are something that need to be looked critically upon, especially if comparing between studies. If not proper accounting for false positives, these number can become distorted. Also, we note that the use-

fulness of databases for comparing and evaluating the results may suffer from inaccurate reports containing many false positives. If an artifact found in one study is submitted to a database of known somatic mutations, variants found at the same position in other studies are more likely to be judged as true, which can make false findings at error-prone sites to accumulate. Controlling for false positives is hence important also for preserving valuable scientific resources.

In paper II an III, we searched for somatic mutations in AML patients to be used as biomarkers in personalised MRD analysis. Such candidates were found in nearly all examined cases. In paper III, 17 of the found SNV:s were tested in deep sequencing and of those 15 was deemed appropriate based on having a low position-specific error rate. All 15 mutations that were utilised in patient samples were shown to be useful for quantification of remaining cancer cells. This suggests that the method to choose appropriate markers from exome-sequencing data was well-adjusted. A statistical model for quantification of variant alleles in deep sequencing data was developed and tested in MRD analysis. The level of detection was determined to 0.02%, meaning that the proposed method meets the requirements for a useful MRD method (Schuurhuis et al., 2018). A comparison to standard MFC-MRD analysis showed that the deep sequencing method had superior sensitivity. In Paper III, detecting the variant alleles in blood instead of bone marrow was also tested with good results. In blood, the levels of tumor cells may be substantially lower than in bone marrow and hence even more sensitive methods are needed. However, to be useful, a future method for regular monitoring of MRD levels after remission in order to detect relapses early, need to be performed based on blood samples due to the invasive procedure when taking a bone marrow sample. We participate in an ongoing study where this is further evaluated, showing promising results for detecting cancer cells with our proposed method in blood samples several months before a relapse is evident. These findings have therefore the potential to improve the diagnostics of AML, both in relation to making MRD analysis possible for all patients and to replace invasive biopsies by blood samples.

In Paper IV, clustering based on miRNA gene expression profiles divided the siNET tumors into subgroups. Thanks to a well characterised patient cohort, the two groups with metastatic tumors could be shown to be associated to clinically relevant parameters such as overall survival and proliferation rate. Also, miR-375 was identified as a potential biomarker for survival based on the miRNA gene expression arrays. This was validated in an independent cohort of siNETs, using tissue microarrays, which showed association to survival time in liver metastases. The supervised and unsupervised methods applied to analyse this dataset were thus able to separate the biological signals from the large levels of noise. Both the finding of clinical relevant molecular subgroups and a potential biomarker for survival might aid in determining prognosis in

individual patients, something that today is hard.

In Paper V, the focus shifted from finding mutations in individuals to instead detecting differences on the nucleotide level between microbial communities encountering different experimental conditions. Each condition is typically sampled a number of times, to be able to take the often high biological variability into account, and hence groups of metagenomes are compared. We developed a statistical model that takes both within- and between-sample variability into account and applies a shrinkage approach for more accurate variance estimation. The proposed model was shown to have a good ability to detect differences in nucleotide compositions, also in cases of high between-sample variability, few samples and small differences. The performance was considerable better than for a standard χ^2-test. Thus, by using tailored statistical methods mutations in microbial communities can be found, even when they only exist in low proportions. This opens up for studying the mutational profile of metagenomes, which is important in, for example, the urgent field of antibiotic resistance.

Analysis of high-throughput data from complex biological datasets is highly dependent on detailed understanding of the experimental design, data generation and the underlying hypotheses. Appropriate methods need to be chosen with care and - almost always - adopted to the specific application. In addition, potential biases and systematic errors, such as differences in experimental setups between samples, needs to be taken into consideration. It should be emphasised that, as a statistician, you need to have an understanding of the field associated with the research question in order to select the most appropriate methodology. It is therefore important with close and efficient collaborations to facilitate knowledge transfer between scientific disciplines. This has been the case for the research reported in this thesis, which has been conducted in interdisciplinary teams including competences in statistics, bioinformatics, molecular biology and medicine. I am certain this has significantly improved the quality and impact of our results.

The experimental methods continue to evolve, with a higher throughput, better quality and the ability to generate new types of information. Still, the fundamental challenges in life science research addressed in this thesis will remain. Life rely on complex schemes of structures and processes, and is truly high-dimensional in its nature. To deepen our understanding of these systems, the use of statistical methods to account for the biological variability will continue to be crucial. Also, implementing the knowledge we gain, and advances in experimental techniques, into our health care and other areas of society is an important task and will remain dependent on interdisciplinary efforts.

References

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18.

Armaghany, T., Wilson, J. D., Chu, Q., and Mills, G. (2012). Genetic Alterations in Colorectal Cancer. *Gastrointestinal Cancer Research : GCR*, 5(1):19–27.

Bagel, S., Hüllen, V., Wiedemann, B., and Heisig, P. (1999). Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli. *Antimicrobial Agents and Chemotherapy*, 43(4):868–875.

Bartram, J., Mountjoy, E., Brooks, T., Hancock, J., Williamson, H., Wright, G., Moppett, J., Goulden, N., and Hubank, M. (2016). Accurate Sample Assignment in a Multiplexed, Ultrasensitive, High-Throughput Sequencing Assay for Minimal Residual Disease. *The Journal of molecular diagnostics: JMD*, 18(4):494–506.

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews. Microbiology*, 13(1):42–51.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–193.

Breveglieri, G., D'Aversa, E., Finotti, A., and Borgatti, M. (2019). Non-invasive Prenatal Testing Using Fetal DNA. *Molecular Diagnosis & Therapy*, 23(2):291–299.

Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E., and Bishop, J. M. (1984). Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science (New York, N.Y.)*, 224(4653):1121–1124.

Cai, L., Yuan, W., Zhang, Z., He, L., and Chou, K.-C. (2016). In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, 6:36540.

Chaudhary, N. and Wesemann, D. R. (2018). Analyzing Immunoglobulin Repertoires. *Frontiers in Immunology*, 9:462.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

Do, H. and Dobrovic, A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clinical Chemistry*, 61(1):64–71.

Fiala, C. and Diamandis, E. P. (2018). Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. *BMC medicine*, 16(1):166.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., and Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, 1.

Gao, Q., Liang, W.-W., Foltz, S. M., Mutharasu, G., Jayasinghe, R. G., Cao, S., Liao, W.-W., Reynolds, S. M., Wyczalkowski, M. A., Yao, L., Yu, L., Sun, S. Q., Fusion Analysis Working Group, Cancer Genome Atlas Research Network, Chen, K., Lazar, A. J., Fields, R. C., Wendl, M. C., Van Tine, B. A., Vij, R., Chen, F., Nykter, M., Shmulevich, I., and Ding, L. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Reports*, 23(1):227–238.e3.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu,

Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H., Stevenson, J. A., Barthorpe, S., Lutz, S. R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J. L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H. G., Chang, J. W., Baselga, J., Stamenkovic, I., Engelman, J. A., Sharma, S. V., Delattre, O., Saez-Rodriguez, J., Gray, N. S., Settleman, J., Futreal, P. A., Haber, D. A., Stratton, M. R., Ramaswamy, S., McDermott, U., and Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3:811.

Goldstein, R. E., O'Neill, J. A., Holcomb, G. W., Morgan, W. M., Neblett, W. W., Oates, J. A., Brown, N., Nadeau, J., Smith, B., Page, D. L., Abumrad, N. N., and Scott, H. W. (1999). Clinical experience over 48 years with pheochromocytoma. *Annals of Surgery*, 229(6):755–764; discussion 764–766.

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., Sverrisson, J. T., Gretarsdottir, S., Walters, G. B., Rafnar, T., Thjodleifsson, B., Bjornsson, E. S., Olafsson, S., Thorarinsdottir, H., Steingrimsdottir, T., Gudmundsdottir, T. S., Theodors, A., Jonasson, J. G., Sigurdsson, A., Bjornsdottir, G., Jonsson, J. J., Thorarensen, O., Ludvigsson, P., Gudbjartsson, H., Eyjolfsson, G. I., Sigurdardottir, O., Olafsson, I., Arnar, D. O., Magnusson, O. T., Kong, A., Masson, G., Thorsteinsdottir, U., Helgason, A., Sulem, P., and Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Haraldsdottir, S., Rafnar, T., Frankel, W. L., Einarsdottir, S., Sigurdsson, A., Hampel, H., Snaebjornsson, P., Masson, G., Weng, D., Arngrimsson, R., Kehr, B., Yilmaz, A., Haraldsson, S., Sulem, P., Stefansson, T., Shields, P. G., Sigurdsson, F., Bekaii-Saab, T., Moller, P. H., Steinarsdottir, M., Alexiusdottir, K., Hitchins, M., Pritchard, C. C., de la Chapelle, A., Jonasson, J. G., Goldberg, R. M., and Stefansson, K. (2017). Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2. *Nature Communications*, 8:14755.

Johnning, A., Kristiansson, E., Angelin, M., Marathe, N., Shouche, Y. S., Johansson, A., and Larsson, D. G. J. (2015). Quinolone resistance mutations in the faecal microbiota of Swedish travellers to India. *BMC Microbiology*, 15.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., Ward, L. D., Arnadottir, G. A., Helgason, E. A., Helgason, H., Gylfason, A., Jonasdottir, A., Jonasdottir, A., Rafnar, T., Besenbacher, S., Frigge, M. L., Stacey, S. N., Magnusson, O. T., Thorsteinsdottir, U., Masson, G., Kong, A., Halldorsson, B. V., Helgason, A., Gudbjartsson, D. F., and Stefansson, K. (2017). Whole genome characterization of sequence diversity of 15,220 Icelanders. *Scientific Data*, 4:170115.

Katagiri, F. and Glazebrook, J. (2009). Overview of mRNA expression profiling using DNA microarrays. *Current Protocols in Molecular Biology*, Chapter 22:Unit 22.4.

King, M.-C., Marks, J. H., Mandell, J. B., and New York Breast Cancer Study Group (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science (New York, N.Y.)*, 302(5645):643–646.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–823.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576.

Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., Zhang, S., and Li, S. (2016). Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PloS One*, 11(1):e0146638.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology*, 79(17):5112–5120.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, 2015(11):951–969.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J.,

Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, 28(3):311–317.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595.

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483.

Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation*, 34(9):E2393–2402.

Lo, Y. M., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., and Wainscoat, J. S. (1997). Presence of fetal DNA in maternal plasma and serum. *Lancet (London, England)*, 350(9076):485–487.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science (New York, N.Y.)*, 349(6255):1483–1489.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1):31–46.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 12(11):R112.

Nilsson, B., Nilsson, O., and Ahlman, H. (2009). Treatment of gastrointestinal stromal tumours: imatinib, sunitinib – and then? *Expert Opinion on Investigational Drugs*, 18(4):457–468.

Oddsson, A., Sulem, P., Helgason, H., Edvardsson, V. O., Thorleifsson, G., Sveinbjörnsson, G., Haraldsdottir, E., Eyjolfsson, G. I., Sigurdardottir, O., Olafsson, I., Masson, G., Holm, H., Gudbjartsson, D. F., Thorsteinsdottir, U., Indridason, O. S., Palsson, R., and Stefansson, K. (2015). Common and rare variants associated with kidney stones and biochemical traits. *Nature Communications*, 6:7975.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics*, 6.

Ommen, H. B. (2016). Monitoring minimal residual disease in acute myeloid leukaemia: a review of the current evolving strategies. *Therapeutic Advances in Hematology*, 7(1):3–16.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A., and Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196.

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, 6(1):5.

Ravandi, F., Walter, R. B., and Freeman, S. D. (2018). Evaluating measurable residual disease in acute myeloid leukemia. *Blood Advances*, 2(11):1356–1366.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics (Oxford, England)*, 23(20):2700–2707.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, 28(14):1811–1817.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC bioinformatics*, 17:125.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

Schuurhuis, G. J., Heuser, M., Freeman, S., Béné, M.-C., Buccisano, F., Cloos, J., Grimwade, D., Haferlach, T., Hills, R. K., Hourigan, C. S., Jorgensen, J. L., Kern, W., Lacombe, F., Maurillo, L., Preudhomme, C., van der Reijden, B. A., Thiede, C., Venditti, A., Vyas, P., Wood, B. L., Walter, R. B., Döhner, K., Roboz, G. J., and Ossenkoppele, G. J. (2018). Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood*, 131(12):1275–1291.

Sengupta, S., Gulukota, K., Zhu, Y., Ober, C., Naughton, K., Wentworth-Sheilds, W., and Ji, Y. (2016). Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic Acids Research*, 44(3):e25.

Shen, Y., Zhu, Y.-M., Fan, X., Shi, J.-Y., Wang, Q.-R., Yan, X.-J., Gu, Z.-H., Wang, Y.-Y., Chen, B., Jiang, C.-L., Yan, H., Chen, F.-F., Chen, H.-M., Chen, Z., Jin, J., and Chen, S.-J. (2011). Gene mutation patterns and their prognostic impact in a cohort of 1185 patients with acute myeloid leukemia. *Blood*, 118(20):5593–5603.

Shin, H.-T., Choi, Y.-L., Yun, J. W., Kim, N. K. D., Kim, S.-Y., Jeon, H. J., Nam, J.-Y., Lee, C., Ryu, D., Kim, S. C., Park, K., Lee, E., Bae, J. S., Son, D. S., Joung, J.-G., Lee, J., Kim, S. T., Ahn, M.-J., Lee, S.-H., Ahn, J. S.,

Lee, W. Y., Oh, B. Y., Park, Y. H., Lee, J. E., Lee, K. H., Kim, H. C., Kim, K.-M., Im, Y.-H., Park, K., Park, P. J., and Park, W.-Y. (2017). Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature Communications*, 8(1):1377.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3.

Spinella, J.-F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., Healy, J., and Sinnett, D. (2016). SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC genomics*, 17(1):912.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.

Thorven, M., Grahn, A., Hedlund, K.-O., Johansson, H., Wahlfrid, C., Larson, G., and Svensson, L. (2005). A homozygous nonsense mutation (428g–>A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *Journal of Virology*, 79(24):15351–15355.

Treangen, T. J. and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1):36–46.

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereau, A., Smith, K., Martin, A. R., Sosinsky, A., McDonagh, E. M., Sultana, R., Mueller, M., Smedley, D., Toms, A., Dinh, L., Fowler, T., Bale, M., Hubbard, T., Rendon, A., Hill, S., Caulfield, M. J., and 100000 Genomes Project (2018). The 100000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ (Clinical research ed.)*, 361:k1687.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43:11.10.1–33.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R.,

and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–1558.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24.

Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., and Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics*, 15:244.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227.

Yost, S. E., Smith, E. N., Schwab, R. B., Bao, L., Jung, H., Wang, X., Voest, E., Pierce, J. P., Messer, K., Parker, B. A., Harismendy, O., and Frazer, K. A. (2012). Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Research*, 40(14):e107.

Zahreddine, H. and Borden, K. L. B. (2013). Mechanisms and insights into drug resistance in cancer. *Frontiers in Pharmacology*, 4:28.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, 30(5):614–620.